Textflo

Version 3.8.1

[User Guide]

Kieran Greer,
Email: help@distributedcomputingsystems.co.uk.
http://distributedcomputingsystems.co.uk/textfilter.html

Table of Contents

1 Introduction

This guide describes a text file processing program that can filter or format text-based content. The application also includes more advanced organiser and search capabilities and might more correctly be termed a text management system. It started as an application that applied basic mathematical operations to text documents, to filter or change the content, but has developed more into the maintenance and use of existing content. The search and organisation capabilities are now quite advanced and can be used to organise or even schedule, your local or online documents. With the large number of documents stored on your computer and online links that you might use, this is a helpful application that allows you to navigate the environment more easily.

For text processing itself, the application can read Text or XML files and can apply a sequence of operations on the text to transform it into a different format or structure. The transformation can involve removing or changing the text, as specified by a filter procedure that can even be saved and re-used. Some operations allow ordered lists to be extracted from files of arbitrary text. The program can also parse and filter the contents of PDF or even HTML files. It can also perform some of the more common formatting operations. Most of the fields that you enter can then be searched over. A separate Organiser application allows you to categorise your local documents or online links and even set deadlines or reminders. A query form allows complex query operations over the content, while a grid format allows for more complex sorts over tabular data. There are also a number of analysis algorithms to help with categorisation, or just understanding the content better.

The key features are as follows:

1. An Organiser application allows you to store your online links or local files, into ordered books and categories that can also be searched over. This allows for querying most of the available information, including keywords, descriptions, notes and content.
2. A Bookmarks form displays a list of ordered file references or links, for any type of file. You can open one of your commonly viewed files or links through a single click.
3. Basic search from a toolbar, or more complex search operations from a Query form. The search facilities allow you to find information from different views and query types.
4. The ability to read text, XML, or parse the contents of PDF or online content (HTML), or binary (Microsoft Office) files.
5. A set of filter operations that can be applied in sequences, saved and retrieved.
6. A set of folders can be used as default locations for all of your related data. The application runs locally only, there is no requirement for a remote server.
7. A grid or tabular format allows you to view database queries, or for operations over specific columns or cells. A log file might have a standard format that can be queried, for example. Conversion to CSV, for example.
8. An analysis panel allows you to select single or groups of files and compare the content based on known clustering algorithms. Also some basic statistical counts.

The main Textflo application is relatively easy to use and works by allowing the user to create a sequence of filtering options that can then be applied to the text. There is some XML (re)formatting, but applications exist that can do this much better, so only a limited amount of XML formatting is available. When reading a file, the application loads it into memory and also reads it one line at a time. You may have to wait a few seconds for larger files to load in.

1.1 Installing the Application

The application is provided as a self-installing executable. Just run the installer and follow the instructions to install the application into the desired directory. You can then access it through the start menu or desktop shortcut. The installer also creates a folder in your root user directory. This folder is called `tffData` and contains important config files and information. You should not delete or move it, or change any of the files in it. You can however add new files and use it as your data repository. Any re-installation will only delete the files that are added as part of the installation. Your own files should not be changed.

1.2 Upgrading the Application

The application comes as an executable program. There is now a very basic licencing system in place for the professional version. After purchasing, you will be sent a licence key that will allow you to use the application for the licence length of time. You simply need to copy the key into your root `tffData` folder in your root user directory. The program should then read it from there.

1.3 Demo Version

The demo version has full functionality apart from some additional features.

1.4 Professional Version

The professional version is the same as the demo version, apart from the following additions.

- **Database:** The demo version is limited to 1000 lines in the database, the professional version is unlimited.
- **Analysis:** Additional analysis options.

2 Application GUI

On startup, the menu shown in Figure 1 opens, allowing you to select one of three options. The first blue circular button is for the main GUI application. The second button is for the Organiser application. The third button is for the Bookmarks form. The Bookmarks are simply a list that can be ordered and selected from. They will open up any document that your default applications allow. The Organiser is quite intuitive and when you start to add file or online links, you will find the search and browse facilities very useful. The main Textflo application is for more specific operations, but still easy to use and you can use the other applications without knowing too much about the main one.

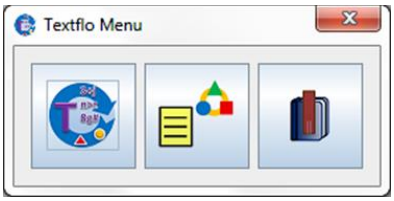


Figure 1. Startup Menu.

The Organiser is also accessible from the main GUI, but more often it is convenient to use it by itself, without interaction with the main GUI. In that case, a few of the Organiser options are disabled. The Organiser is described in a separate document, while the main application is described in the following sections. The Bookmarks form can also be opened from the main GUI, with a config option to allow this automatically, when the main GUI starts-up. Again, that is just a preference.

The main Textflo application consists of three different panels that perform different kinds of operations. Two of the panels can filter or format the text, while the other performs a limited amount of analysis over the text. The Figure 2 graphic shows what the GUI looks like. This also shows a file describing a food menu, loaded into the application in its original format. The panels in the GUI are as follows:

- The `General` panel allows for general filtering operations over the whole text file.
- The `Manual` panel allows the user to manually specify certain cells or areas to filter.
- The `Analysis` panel can perform some statistical analysis of the text.

The function of each panel is described in the following sections. To illustrate potential usefulness, you can see in the figure that a text area has been selected, or highlighted. It is possible to select more than one area, either manually, or through one of the query evaluations. All of the distinct areas can then be filtered, independently of the whole

document, with the result placed back into the whole document again, for example. Note that if you click on the text area, that will automatically highlight the row of text.

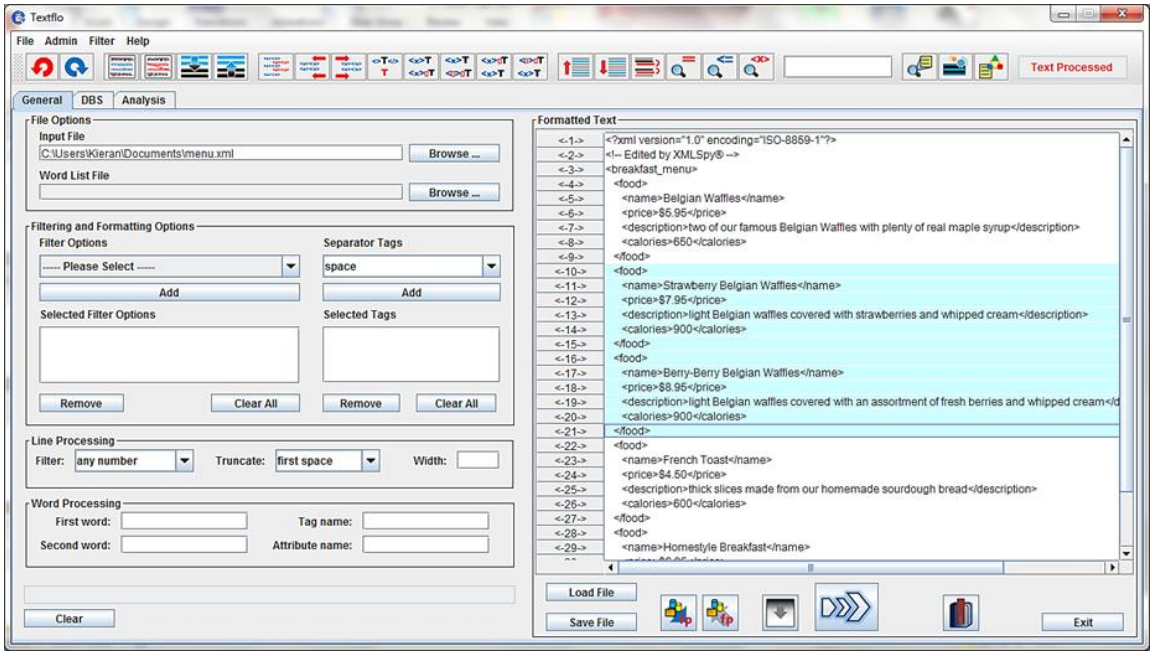


Figure 2. Main GUI Application

2.1 File Types

While the file browser should allow you to open any type of file, the following types specifically are processed by the application.

2.1.1 Standard file types

- **.txt:** these are plain text files, which is essentially what the application processes.
- **.xml:** these are XML files that can be parsed and validated as XML.
- **.pdf:** these are PDF files that are converted into plain text before processing.
- **.doc, .docx, .ppt, .pptx, .xls, xlsx:** these are Microsoft Word, PowerPoint or Excel files that are converted into plain text before processing.
- **.html:** these are html file that are saved on your computer. You might typically try to retrieve an html file from the internet instead.

2.1.2 TFF-specific file types

- **.fpr:** these are stored filter procedure files. When the application saves a filter procedure to a file, it adds this extension to the file.

- **.anls:** these are analysis files created by the application. When the application saves its analysis, it adds this extension to the file.

2.2 Menu Options

There are a number of menu options as follows:

2.2.1 File Menu

This contains options for loading or saving files. The following options are available:

- **Open File:** this option allows you load a data file in, in one go. This is instead of browsing to the file and then clicking the `Load` button
- **Save File:** this option allows you to save the filtered or re-formatted text to a file.
- **Save File As …:** if the file field is empty, you cannot save the file directly, so you can use this option to browse to one and then save to it, the current text contents, in one go.
- **Clear All:** this option clears all of the current entries and the processed text.
- **Save Analysis Config:** this option allows you to save a configuration file. This is an analysis configuration files that defines what analysis options are carried out.
- **Save Analysis File:** this option allows you to save an analysis file. This is an analysis of the selected text file.
- **Save Analysis Comparison:** this option allows you to save an analysis comparison file. This is a description of the similarity comparison between a number of analysis files.
- **Exit:** this option terminates the application.

2.2.2 Admin

Some very basic admin or configuration is possible:

- **GUI Config:** automatic configuration of the GUI will be added here. This opens a form, where currently the only option is to automatically open the bookmarks when the main GUI opens.
- **Open at Lib (Win OS):** this is a convenience option that probably only works on the Windows OS. It will open Windows Explorer at the default folder location, so that you can easily access the files there. This should help if you need to delete or change any of them.

2.2.3 Filter

This contains options to help you to filter your documents or text. The following are available:

- **Load Filter:** this option allows you to load a stored filter procedure into the GUI. The values in the form are displayed in the main panel boxes.

- **Save Filter:** this option allows you to save the current filter settings to a file as a stored procedure. These can then be re-loaded to allow you to quickly set up a particular filtering operation.
- **Query:** this option open a query form, to allow you to execute search queries over your text document.
- **Organiser:** this option opens an organiser form, to allow you to organise or group your documents based on their content.
- **Temp File Analysis:** this option can display a set of recent files or references that you have looked at, opened, added, or whatever. It can also suggest lines for separate processing, for example, lines that repeatedly occur.

2.2.4 Help

This contains options for displaying help or checking you applications version. The following options are available:

- **Online User Guide:** this option allows you load the online user guide into your browser for viewing.
- **Check for Updates:** this option allows you to check that your application version is the most recent. It compares you application's version number with the one specified on the web site. If there is a difference, then a message informs you of that.
- **About:** this opens an about box with some general information.

2.3 Toolbar

The application also comes with a toolbar for quick access to certain formatting options. A summary of these with their related button are shown next.

This button undo's the last operation, up to 5 of the previous operations.

This button redo's the last operation, up to 5 of the previous operations.

This button keeps only the highlighted text areas.

This button removes only the highlighted text areas.

This button removes all of the highlighting, and associated indexes.

This button re-highlights any text as specified by the stored indexes.

This button performs a pretty format on an XML document.

This button removes one indentation from the selected text.

This button adds one indentation to the selected text.

This button removes the XML tags from the selected text but keeps the content.

This button separates the XML tags from the content.

This button separates the XML tags from the content and also the element names from the element brackets.

This button re-joins the XML tags with the content.

This button re-joins the XML tags with the content and also the element names with the element brackets.

This button works with the search options to automatically move the previously highlighted section to the top of the text output display. If you right-click this button, a small form opens that allows you to set a larger jump size. The number relates to the number of highlights that are skipped, not the number of lines.

This button works with the search options to automatically move the next highlighted section to the top of the text output display. If you right-click this button, a small form opens that allows you to set a larger jump size. The number relates to the number of highlights that are skipped, not the number of lines.

This button works with the search options to automatically move all selected lines down one level.

This button performs a search for lines with exact words, in the whole document.

This button performs a search for lines that contain, in the whole document. Alternatively, you can type the search term into the text area and then press the `Enter` key on your keyboard to start the same search process.

This button highlights XML sections that are contained inside of elements with the specified tag name.

This button allows for a more sophisticated search, with different comparison options, to select the lines to highlight. This also filters the existing highlighted text.

This button opens a form that performs a very simplistic line comparison and suggests lines that occur more frequently. It might be useful for determining what header or footer lines are present, for example. You can then highlight or delete them.

This button opens the organiser form that can be used to group or organise document references through categories and keyword lists.

- The search options now have a text field also in the toolbar, where you enter the text sequence to search for.
- There is also a status field that will let you know when the text is being processed and when the processing has completed. This is helpful during longer operations.

2.4 Filter Buttons

As well as the toolbar, there is a group of buttons at the bottom of the form. These can be used for the following:

This allows you to browse to a stored filter procedure, to load in the details.

This allows you to save the current filter settings as a stored filter procedure.

This opens the HyperSQL database manager interface, allowing you to connect to your database.

This clears the form of all of the current filter settings.

This opens the bookmarks form.

This executes the current filter.

3 General Panel

This panel can perform filtering and formatting processes over the whole text document and is shown in Figure 2. You specify a number of filtering/formatting options that are to be performed in sequence and then run them to change the text. This panel consists of a left-hand side with the filtering/formatting options and a right-hand side that shows the document text. The right-hand side also contains a set of buttons to run the filtering or formatting processes.

3.1 Format Buttons

You select the filtering options from the left-hand side panel, but the filtering is performed only when you press the large button in the middle of the bottom right button panel. The button panel also allows you to load the text file into the GUI, save newly processed text, or open and save filter procedures. These are described in section 3.4. The following processing options are available from the bottom right button panel:

- If you press the `Load File` button, you are able to load in the text of the selected file path. A dialog box opens and asks you if you want to load it in as a text file or an XML file. If you need to perform some XML-related processing, then you can check that the file can be properly read in XML format.
- The `Save File` option should then be used at the end of the filtering process, to save the newly created document.
- The `Load Filter` button allows you to load a stored filter procedure into the GUI, to quickly setup a filtering operation.
- The `Save Filter` button allows you to save the currently selected filtering options to a file, as a stored procedure. This will only save filter details and not actual text content, but it can then be re-loaded to allow the filtering operation to be run again.
- The button to execute a process now shows a large arrow icon image. You click this to start any filtering or formatting process. This is always performed on the text currently loaded into the `Output` text area. If the output display is empty, then you are asked if you want to load in the specified document first. If you want to restart the filtering process on the original document, you need to `Load` it in again first.
- The `Clear` button removes the current filter settings. The same main menu option also deletes the text and related file path.
- The `Bookmarks` button open a bookmarks form with a single ordered list of links.
- The `Exit` button exits the application.

3.2 Bookmarks

Bookmarks are a new feature that display a single ordered list of links. These links are permanent and do not change unless the user manually changes them. They are for convenience, as you can find your most important documents in only one or two clicks, without any additional search process. You can also configure the main GUI to open this form automatically. If, for example, you are reading a paper, then you might want to view it properly in the default application viewer, before deciding to filter it in the Textflo application. To add a new bookmark, you can drag the file from your system folder to the bookmarks button, or use the config form as described next. Figure 3 shows what the bookmarks form looks like.

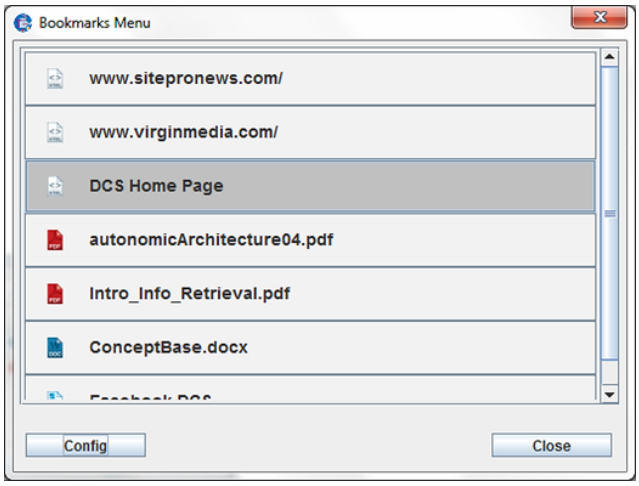


Figure 3. Bookmarks form.

From the list, you simply click on one of the links, to ask the default application of your OS to open it. The file types are varied, for example, there are two HTML files, one PDF and one docx file, in the figure. Bookmarks can be specifically ordered, or placed in a more general ordering. Adding bookmarks can be done by dragging files to the button on the main GUI. It can also be performed through the bookmarks config form, shown in Figure 4.

You can again drag a file, this time to the top `File URL` text field; or you can manually enter any file path there and then click the `Add` button. You can also add a bookmark from the `Organiser` application, when the related name can be displayed instead. Bookmarks can be given a specific index value for display, or can be ordered more generally. At first, they get added to the general list, shown in the combo box. The bookmark that is currently selected is then shown in the bottom text field that cannot be edited. If you click the lower `Add` button, the currently selected bookmark will be added to the list just above. Once on the list, it can be selected again and moved up or down, using the arrow buttons. The bookmarks on the list are always ordered first and in the specific ordering that is specified. If you select

the bottom bookmark on the list and move it down, it gets removed and added to the general list. The currently selected bookmark can then be removed, using the `Remove` button. Note that the currently selected bookmark can be from either list.

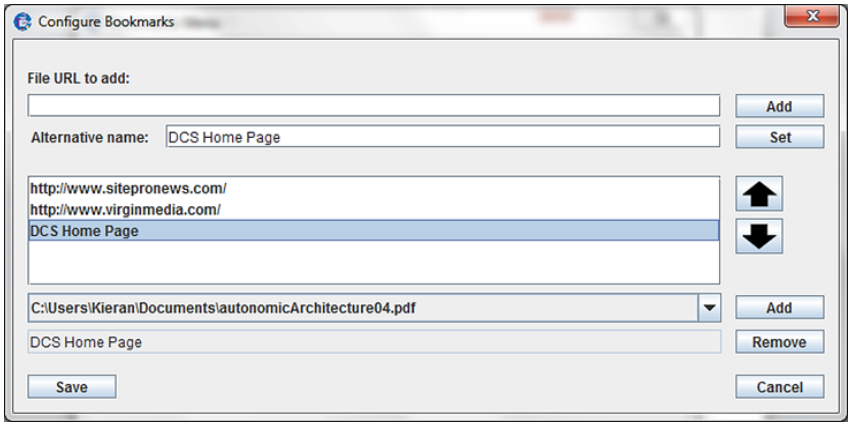


Figure 4. Bookmarks configuration form.

There is also an alternative name field. If you enter a value there and click the `Set` button, that name will be displayed for the currently selected bookmark. The bookmarks are saved in a separate file, so this will not affect any Organiser books and a backup file is also created. While the view may update, it is important to save the new ordering before exiting, for it to be made permanent.

3.3 Formatting Settings

To perform any filtering or formatting, you need to specify what operations should be performed. These can be specified in the left-hand side group boxes. Note at the bottom of the LHS, there is now a single text field, where information can be output. The `Clear` button will then automatically remove it. The information might be, for example, the number of lines that have been selected (and highlighted) from a search.

3.3.1 File Selection

In the `File Options` group box, you can select the file to format or filter. This box is now editable, so you can enter an http web address and load a file from the internet instead. The file can be loaded into the GUI using the `Load` button. This is also a useful operation simply for checking that the contents are valid. If you do not load the file, then the first processing operation will load in the file contents. You can either load in the file as lines of text or, if it is an XML file, you can format the file into XML. An XML file can also be read simply as text, but will lose its nested elements indentation. Figure 5 shows the window that lets you

make this choice. Note however that the process can be quite slow and the GUI will be unresponsive when loading in the data. It would be best to try a smaller file first.

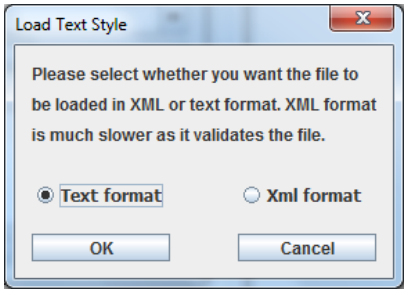


Figure 5. Load file as box.

If you load the file into the GUI, reading it in text format is much quicker and should therefore be recommended for larger files. Loading in XML format also checks that the file can be read as a valid XML file and so is much slower. There is also a menu option to open a file and when selected the file is also automatically read and displayed in the GUI. Alternatively, you can browse to a file path and perform processing operations without first loading the text into the GUI. This should be quicker, because the GUI components are then not updated with the text content first. You can also load in a PDF file, when the text content will be parsed and displayed. It might not be possible to convert all PDF files, where the operation assumes that the file has no security or other restrictions and can be converted as is.

3.3.2 Word List Filter Selection

You can select a file that stores words to be used for filtering options. This `Word List` file is a file that needs to be created before running the application. This file should have a single word or term on each line, where these terms will be used to filter the text when the appropriate filtering option is selected. The file `commonWords.txt` included in the example files folder is an example of this sort of file. If this file is specified, it is automatically loaded and used. It also overrides any words that are manually entered into the other filter fields. A message should be output however to make this clear. Because of that, the text field it is now editable, where you can also remove any browsed file path again, to allow manual entries to be used instead.

3.3.3 Word Ordering File Selection

There is also an option to load in a word ordering, to be used to sort a filtered set of words. This is now part of the DBS panel. This word list is then displayed in total in the related list, from where it can also be changed manually. The list sorts are now part of the DBS panel and this specific word list relates to them only.

3.3.4 Start Formatting / Reformatting

To start a formatting or filtering process, you firstly need to select a number of filter options. This is described in section 4. The text is processed in the order of these options. If you select a processing option from the list, the labels of the related data entry fields will be highlighted, which should help with data input. To perform some filtering you need to click the large arrowed button. The formatted text will then be displayed in the `Output` text area. You can then clear the selected options and repeat the process for each filtering/formatting operation. If you want to filter the original file again, then you should click the `Load File` button to load in the original text first.

3.4 Stored Filter Procedures

Stored filter procedures are filter/format procedures that can be defined through the `General` panel and then saved to a file. They are essentially a copy of the filter/format description displayed in this panel. They also really tie the whole application together, as you can test filtering operations and then save useful ones to a file. You can then load in any text file, reload the filtering operation and apply it to the text in one easy step.

3.5 Popup Menu

The general text table includes a popup menu to allow a row of text to be edited. Any text changes are also copied to the underlying text model and so they need to be made specifically. Note that the cells need to be highlighted first before they can be selected. The popup menu currently has the following options:

1. **Open:** If the text row is a valid link or reference, it can be opened using the system default application. If it is just a line of text, then a message will be displayed instead. You can list file references in the main text if you perform a folder search (see section 5.2.3.1), for example. If a line is selected and not a file path, then the main file path will be loaded instead. So you can view the original document in the default viewer using this menu option as well. This also occurs if a line is not selected.
2. **Edit Row:** This opens a window with the text of the selected row included. You can then either edit the text or cancel the operation.
3. **Copy to Clipboard:** This copies the currently selected text section to the system clipboard. If you have performed a search, for example, you then need to manually re-highlight the section of text that you want to copy. This will necessarily remove the other highlighting, but it is necessary, to let the program know what exactly you want to copy.
4. **Go To Line:** This does not require a line to be selected first and will scroll the text to the entered line number.

4 Filter and Format Options

To start a formatting or filtering process, you firstly need to select a number of options. This is done in the `Filtering and Formatting Options` group box. The currently available filter options are displayed in the top left combo box. You need to select at least one filter option to perform any filtering/formatting. You can select several options in sequence, when they will then be performed automatically on the text in that order. If you select a processing option from the list, the labels of the allowed data entry fields will be highlighted.

After you select an option, the list will try to update itself so that only the remaining appropriate options are left, although this is quite open now and so most options are generally available. If you then remove an option, this can change the list back again to what it was. However, if you perform a sequence of filtering and then clear the selection, you will be presented with the complete list again to choose from. So this process is more of a guide than a hard rule. You will have to determine for yourself what options are appropriate and will work in sequence. The list of options has been categorised to try and show what each option relates to. The options themselves will be described next under the same categories, where Appendix A of section 12 gives a more specific summary of each option.

4.1 Drag and Drop

For convenience, it is possible to select a line of text from the output text window and drag it onto one of the editable text fields. This could help with specifying certain text values or symbols that need to be processed or removed, for example. The whole text line is entered, but it can then be edited, where certain symbols that are difficult to enter manually can be used, for example.

4.2 Basic Formatting

These operations allow you to perform some basic formatting on the text document. Options include upper case/lower case conversion, replacing one word with another, or removing double spaces. One option allows you to reformat the text to a new maximum width. You can reformat the whole text to produce a new single paragraph document where each line is a maximum length of the specified width. The maximum line width is specified in the `Width` text field in the `Line Processing` group box.

There is now also an option to reformat to a specified width, but keep new lines that already exist, from specified non-whitespace separator characters, as well. The separator characters can be defined in the `Separator Tags` list and if any are non-whitespace and form the start of a line that new line is kept no matter what the width of the current line is. So a new line could be created before the specified maximum width, if a separator character is

encountered first. This option can best be used along with the options to create single column lists of words. It is probably better to create a single column list of words based on specified separator characters and then reformat to a new width, but keep the existing new lines as well. The single column lists can create the new lines before or after the specified separator character and keep or remove the separator character as well.

There are also options to allow you to truncate or trim a line. Truncating a line means that the line will be truncated before or after the first occurrence of a specified character or word. By default, you can truncate a line at the first occurrence of a white space, tab, letter or number. You can also then enter your own character or word and that will be used instead. If you have a file of text that has been copied from a table, for example, you may want to keep only the first column of words and remove the other numerical columns from the text. This is the sort of thing that the line truncation can do, where you would truncate at the first numerical digit. You also have the option to include the truncation character or word as part of the final line.

4.2.1 Separators Tags

As well as filter options, you can enter a list of tags in the `Separator Tags` group of components. These can be used to split lines of text on the specified words or characters. There is a default filter option to generate a word list, where the text is separated by whitespace only. There is however also an option to allow you to use other characters as the separators, for example, commas or periods. These will then be used instead of the default whitespace and also removed from the final text. This is the list of separators created in this group of components. To create your own list of separators you can manually enter the character or word into the combo box and click the `Add` button. The combo box also contains default words to represent the whitespace characters. You can also select these and then when the formatting takes place, they will be converted into the appropriate character.

4.3 XML-Based Filtering

There are also a number of formatting/filtering options for XML-based text, although they are centred more round filtering the text than reformatting it into good looking XML. Because of this, a number of options focus on separating the XML tags from the rest of the text and then allowing you to re-join the tags again to form a valid XML document.

4.3.1 Convert XML Tags to or from Words

When you load in an XML document, it will contain XML element tags with names, as well as the content of those elements. This can be difficult to filter as the element content is typically aligned right beside the element tags. Also, the tag names might be important, but they are surrounded by the XML element formatting characters. You have the option to convert the element tags into individual words. This is done by placing spaces between the XML formatting characters (<, </ or >, etc) and the element names, and also between the

formatting characters and the text content. It is then possible to read each element name as a specific and individual word. You could then filter the text based on the element names, for example. An option then exists to convert the words back into XML element tags again. This is only possible if the conversion the other way has been done first, that is, the XML character formatting is still present in the document. So you could convert the elements to words, filter to remove a number of these elements and then convert back into XML. The options from `Remove XML tags` to `Words to Tags` perform the task of separating the tags from the text content and then re-joining them again.

The other options in this category allow you to convert text to element or attribute values. This requires you to enter additional words to be used as a search option or as a tag or attribute name. These values are entered into the Ordering and XML group box as described next. Another option allows you to keep just the text that would be read in an HTML file. An HTML document can be loaded into the application as a text file. You can then select the `Keep HTML Reading Content` option to extract only the text in the document that is to be read. All of the other web page formatting and layout information is removed, including all of the XML tags.

4.3.2 Ordering and XML

This group box has components to allow you to enter values that can be used for user-specified ordering or limited XML formatting.

- The left-hand group of components allow you to enter your own word ordering list that can be used to order the text. You enter a word into the `Word Sort Order` text box and click the `Add` button to add it to the list just below. You can then choose a formatting option that uses this ordering to order the words in the text document.
- The right-hand group of components allow you to enter values for XML tag names, or words to be formatted by the XML formatting process.
  - If you enter a tag name (`Tag name` box), you can then surround each line of text with this tag name, converting it into XML.
  - If you also specify the word value (`Attribute name` box), you can surround only that word with the XML tag. Note however that this will not allow you to create nested XML structures, etc. which an XML Editor would allow you to do.
  - There is also an option to convert the text of an element into an attribute value, by entering the element name (`Tag name` box) and the attribute name (`Attrbute name` box) that the element's content will be converted into.

4.4 Word or Line Processing

This category of options allows you to process individual words or lines, specifically to choose what words or lines to keep or remove. There are also options to remove duplicate

entries. You have options to read the comparison terms from a file, for multiple entries, or to enter a single term manually.

4.4.1 Remove Separators

This allows you to select a number of characters that should be removed from the text. You then need to enter the separator list into the `Separator Tags` area.

4.4.2 Remove or Keep Text

These options allow you to read terms from a file, or a text field and to remove whole lines or the words themselves, from the main text.

4.4.2.1 Remove Terms in a File

You can either keep or remove, lines or specific words that are related to a word list that you create. The path to the word list file should be loaded into the `Word List File` text field. The file `commonWords.txt` included in the download zip is an example of this sort of file. You can then either remove all lines that contain any of the words, or keep only those lines that contain any of the words. To keep or remove lines, you have the choice of selecting lines that 'contain' any of the words or text in the list, or that 'match exactly' the words or text in the list. You can perform a similar action for removing individual words instead of whole lines. Note that the word list file field is editable and so you can clear or delete it easily.

4.4.2.2 Manually Enter a Term

If you do not want to enter a file with a list of words, you can leave the `Word List File` path empty and enter a single word or text sequence into the `Filter` text field, in the `Line Processing` group box. It is a combo box with default values, such as 'any number'. You can however overwrite this be entering any specific text words instead. A file is checked for first. If there is no file to be loaded, this field will be checked and if it contains a word sequence, that sequence will be used to select the lines or words to keep or remove. This field now also has some default options. If loading in a PDF file, for example, a complex document can have lots of formatting, with irregular character sequences. The default 'general' options of `any letters`, `any numbers` and `any symbols`, can be used to remove or keep lines that contain any of these characters in general. So this does not apply to matching an exact text sequence, but to a general character type. In that case, the 'with' option matches to lines that contain the character type only, while the 'contains' option matches to lines that contain the character type along with possibly other character types as well. These general character types only apply to the remove / keep lines options though. Appendix A describes these options further.

4.5 Single Lists

This section allows you to generate lists from the unformatted text. Lists are formed by separating the text on specified characters. This could generate a list of single words, or lists of word sequences that are separated on certain characters, for example, new lines. If you do not enter any characters yourself, the words are separated on whitespace characters. Alternatively, you can specify a set of separator characters in the `Separator Tags` list and these will be used instead. The single column lists can create the new lines before or after the specified separator character. There is now also an option to keep the separator tags as part of the formatted text. If any of the tags are non-whitespace, you can choose to keep them as part the formatted text. The tag will appear as either the first character of the next new line in the text, or the last character on the line before the new line is created.

There is also the option of creating single lists of words from the XML element tag names, instead of from the text content. The text must be in valid XML format that can be read by the XML parser. If you are looking to analyse this further for some sort of pattern, the XML structure itself can then be analysed.

4.6 Producing Sorted Lists

This category allows you to generate sorted lists of words based on certain ordering criteria, which allows for a certain amount of data mining in the text, to see what patterns might exist. You can, for example, order words in decreasing or increasing order of their letter value, or based on a pre-specified word order. In that case, the algorithm looks for the word order in the single list of words and can remove or place any other words at the end, leaving only the list of ordered words in sequence. This might be useful for finding popular sequences of words, or for categorising the text, for example. The descriptions for this section have been moved to the text analysis guide that can be downloaded from the main web site at the address http://distributedcomputingsystems.co.uk/Documents/tffTextAnalysis.pdf. A text analysis guide has been written to reduce the size of this document and to store the more technical information about exactly what the analysis does. The sort options have now been moved to the DBS panel (see section 10) as they would typically be associated more with tabular data and single columns of terms.

5 Query Search Options

This section describes the query search options that are available. The previous sections allowed the filtering process to find lines that contained certain text sequences. It would then remove all other lines and keep only the matching lines. It is also possible to search over the text document to find lines that match a more flexible set of criteria. The main difference with the search options is that when they find any relevant lines, they keep them as part of the whole text document and only highlight them. These lines can then be selected and processed further and separately, if desired. The processing result can then still remain as part of the whole document. The search options need to be performed in isolation of other filtering operations. Because of this, they are included as toolbar options and not included in the main filtering list.

The query options can be broadly divided into ones available from the toolbar and ones available through a new query form. The differences between these are as follows:

- **Toolbar Search:** the toolbar allows you to enter a search term or phrase and search over the text to find matches to it. There is some sense of process, as you can also choose to search only the next lines of currently selected ones. There are also arrow buttons that can move or direct a search to the next area that was found. This is useful if you have no clear idea of what the text is about and you just want to randomly select words or phrases to see if they are contained in the document.
- **Query Form:** for a more complex search, a query form can also be opened from the toolbar. This has been re-designed so that the different types of search (text-based, XML-based, or analysis feedback) are available from different panels. The query language is integrated however and so the query display is always in XML format now.

5.1 Toolbar Search Options

There are two different ways to search over the text. The first option is to enter your search term into the text field on the toolbar and then select one of the toolbar search buttons, also described in section 2.3. Figure 6 shows the toolbar section that is related to the search.

Figure 6. Toolbar buttons for basic search operations.

There first two toolbar buttons that will scroll to the previous or the next highlighted section of text, respectively. This allows you to move to through highlighted sections more easily. The third toolbar button automatically moves all of the selected (or highlighted) lines down to the immediate next line. This is useful for moving through nested sequences of text. You can then query only the selected lines of text as well.

For search query options, the toolbar buttons can look for lines that match the text exactly (=) or contain the text in any sequence (<=).You can also search for XML-specific tag names (<X>). For this, you should include only the tag name and do not include any of the enclosing brackets (< or >) that define an XML element. These searches will search over the whole document. It might be confusing when lines are highlighted or not, so these toolbar options always search the whole document. The toolbar search options are therefore as follows:

- *Find lines with*: this requires a match with a whole word.
- *Find lines that contain*: this requires a match with any text sequence.
- *Find Xml sequences*: this requires a match to an XML element tag name.
- *Perform a more complex query*: this can be used to filter the text based on more complex comparison specifications.

The final toolbar option, on the other side of the text input box, opens the `Query` form that allows you to enter a more complex or flexible set of search criteria. The query form, described in section 5.2, allows you to search over the highlighted lines only. Therefore, a set of selected lines can be filtered through the query form. What you can then do is move all lines down one and execute another query through the form. This will then only look at and select, the lines that came directly after the first set. So if you have some sort of nested sequences, you could search for the top level lines with one search and then refine this with a search for only specific lines the next level down. So with this combination, you can perform an XPath-like query over XML text, for example. In this case, the selection is done manually and one step at a time, but it could have the same sort of result.

5.2 Query Form

The query form has been re-designed around the type of text being searched. The display is a text-based description, rather than the strict XML format that the program uses. It contains the same elements but might be slightly easier to read. At the bottom of the form there is a `Save` button. If allowed, it is enabled and will open at the base Textflo folder, to allow you to save the query as an XML script in one of the folders, probably the files folder. It should be saved as XML, when it can be loaded in again using the ‘Stored queries’ panel. There are now 5 different panels - one for each general query type:

- **Text-based queries:** these are queries that search over whole lines of text. They do not consider any real structure and therefore really only contain a set of constraints that need

to be matched to. They still include the AND/OR-style queries and also the queries that search for regions or areas in the text.

- **XML-based queries:** these are queries that search over XML text specifically. As a result, they can consider structure in the text and contain a pattern section as well as a set of constraints.
- **System queries:** these are queries that search for files or folders anywhere on your computer.
- **Analysis queries:** this panel allows the feedback of analysis results, or less conventional searches. It can use the clustering results of the 'Line Cluster' analysis option, or perform a very general search over folder contents.
- **Stored queries:** allows you to save a constructed query, load it in again and execute it.

5.2.1 Text-Based Queries

This section describes the text-based query options. Figure 7 shows what the `Query Form` with this panel selected, looks like. Text-based queries can take an `AND/OR` format. The AND statements must all be satisfied to allow a line to be included. The OR statements then allow a number of different sets of conditions to be considered. Each OR statement is represented by a new query, while each AND statement is represented by a new condition or constraint in the same query. Words or terms are only considered as they are written, where you can use the `Case` check box to include case sensitivity and the `Wildcards` one to include wildcards. The panel layout and query description also contain some other important elements:

- The `As Text` search type defines a text-based query.
- The `boolean` check box options are available to each constraint separately and not the query as a whole. They are added each time you add a new constraint part.
- The query type can be either `Text` or `Numerical` and both can be included in the same query process.
- Each 'OR' structure is represented by a new query that can contain a completely different set of conditions or constraints. This is created using the `Add` button. To select the different OR queries, there is a box with a query number: q1, q2, etc. This selection determines what query the next constraint is added to. So an `Add` button click will add a new alternative OR query and an `AND` button click will add a new 'constraint' to the currently selected query.
- Another combo box at the far top right shows a value of `All`. This box relates to the new database – sorts grid of the DBS panel (section 10). The 'All' value is the default value. If you wish to query each whole line of text, then you do not change that. If you have loaded the text into the grid structure of the DBS panel, then it gets parsed depending on the separator characters, with different text placed in different columns or cells. This is particularly useful for tabular data. It also means that you can query certain columns only, where you can specify the column number as part of the query constraint. A value of 1,

for example, would mean to execute the query only over the text stored in column 1, and so on. The new DBS panel also allows for complex sorts that are described in section 10.

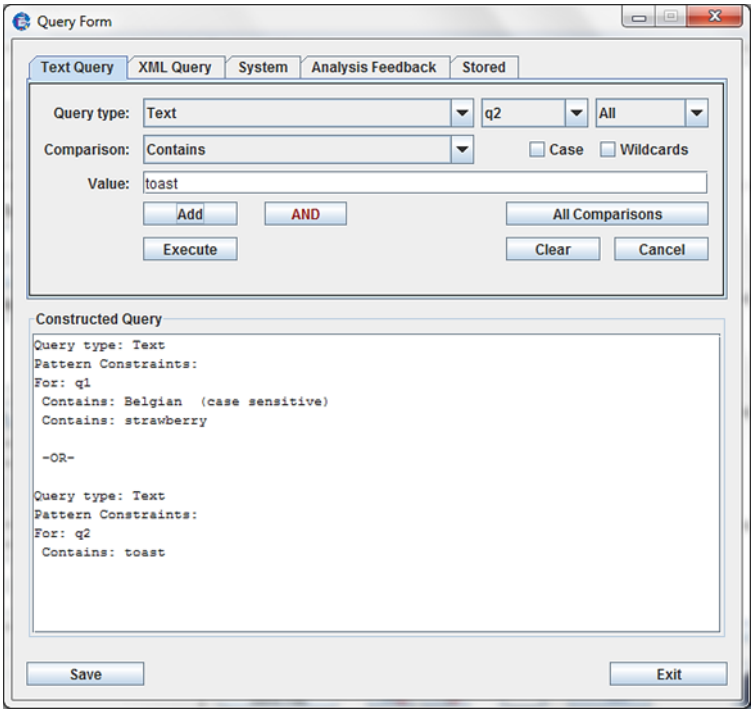


Figure 7. Filter Text-based Query form.

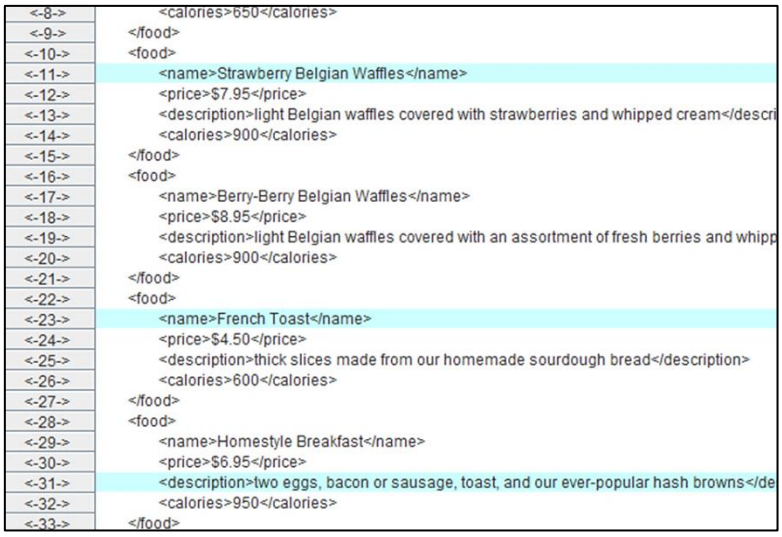


Figure 8. Lines highlighted (selected) by the query.

The query of Figure 7, for example, is stating that for a line to be highlighted, it must contain both the words 'Belgian' and 'strawberry', where Belgian is case sensitive. The list of constraints can be seen below each query – labelled as q1, q2, etc. There is then a second option, where lines that include the word 'toast' can also be accepted. If this query is executed on the whole menu document of Figure 2, then the resulting lines, shown in Figure 8, are highlighted. The query process is a little bit different in the sense that it does not look for a specific variable to evaluate, but evaluates the whole line. With text, it looks for the exact text sequence or one that contains the text sequence.

5.2.1.1 Text-Based Comparison Types

The `Type` field of the query specifies the constraint or comparison type, while the `Value` field specifies the comparison value. There are two distinct types of text-based query:

- **Lines that contain or equal certain criteria:** you can search over lines to check if they contain exactly, or as part of a sequence, the specified search term. These are the `Contains`, `Not Contains`, `Equals`, `Not Equals` options. You can also use `Line starts with`, `Line does not start with` options to check the beginning of a line.
- **Line ranges that start / end with certain criteria:** this allows for a more general type of query that can try to highlight ranges or areas in the text. This search only requires you to enter a 'start' and an 'end' value to search for. Alternatively, you can enter a start term and then a line range, for before and after any lines that match the term. For example, you can select a word or term that a starting line should have (`Start line contains`, `Start line contains exactly`). If lines are found, you can then select a word that an ending line should have (`End line contains`, `End line contains exactly`). The region between the start and end lines that are found will then be highlighted. Alternatively, from the start lines, the options `Lines before selected` or `Lines after selected` allow you to specify line numbers before and after the start lines, where that region is highlighted instead. Values of 2 and 3, for example, would highlight from 2 lines before the found search term to 3 lines after it.

If any of the constraint types are not currently present, you can use the `All` button to reset back to all of the constraint types again.

5.2.1.2 Numerical Comparisons

It is also possible to perform some level of numerical comparisons. This can be used to look for numbers in any position on a line that satisfies the specified constraints. A numerical query is constructed in the same way as a text-based one. If evaluating numbers, then each line will be parsed into tokens separated by spaces and the numerical evaluation applied to each token separately. If the line is an XML element, then the XML tags will also be removed first and only the content parsed.

If a token has non-numerical characters at the start and end, but a number in the middle, the current decision is to trim the non-numerical characters from the start and end and still process the numerical part. This could result in invalid numbers being considered, but it will also help with poorly formatted text, or the removal of punctuation. This process therefore will perform some guessing, but the hope is that it is likely to be correct more often than incorrect. It will also consider negative numbers and try to add a minus sign if it is parsed somewhere valid along the token.

So the following numerical representations are valid or invalid:

<price>$5.50</price>	valid value of 5.50
<price>line with number $5.95</price>	valid value of 5.95
Text line with number 6.5	valid value of 6.5
<calories>6.50abc</calories>	valid value of 6.50
abc650…	valid value of 650
ab1c2de	invalid value

If all lines were to be queried with a numerical evaluation of greater than 6.0, then lines 3 to 5 would be returned. Note that the whole XML element must be on a single line for the tags to be automatically removed. Regarding negative number representations, the following currently applies, where the final parsing example is the most dubious:

123-456	invalid value
-123	valid as -123
-$123.45	valid as -123.45
-abc123	invalid as negative, but valid as 123
-*£%123	valid as -123

5.2.1.3 Wildcard Characters

The text-based query engine can also handle wildcard characters. Note that numerical comparisons do not allow wildcards. To use wildcards, you need to click the `Wildcards` check box. The wildcard specification spans across only one word at a time. The query engine will then take the following characters to be wildcards and not as standard text:

- '*****' if this character is entered, it will represent any number (one or more) of characters until the next exact specification. So, for example, 'B*n' would represent a word starting with 'B' and ending with 'n', with any number of characters in-between. Therefore, Belgian would be included.
- '**?**' if this character is entered it will represent a single character that can be anything. So, for example, 'B?????n' would represent a word starting with 'B', then 5 characters that can be anything, and then an 'n'. Belgian would again be included.

The wildcard characters can also start or end a search term. Also for convenience, the query engine has the option of `starts with` instead of `equals` or `contains`, for text matches. This means that you can search for words that start with something, but can end with anything, and might be easier than entering wildcard characters in some cases.

5.2.2 XML-Based Pattern Search

While the previous option searched for specific lines, the `XML Query` tab allows you to construct a query that will look for XML-based patterns instead. This can therefore also handle X-Path style queries, where the path through nested XML elements can be specified through the element names and associated sets of constraints. An XML-based query can be constructed through specifying a set of element names to search for in sequence and then also specifying a set of constraints on each of the elements. The query structure is saved in XML format and has the 'pattern' section first and then the 'constraints' section. Figure 9 is an example of one of these queries. Not every pattern element needs to have a constraint and it should be possible to miss out elements, but you need to keep the nesting structure.

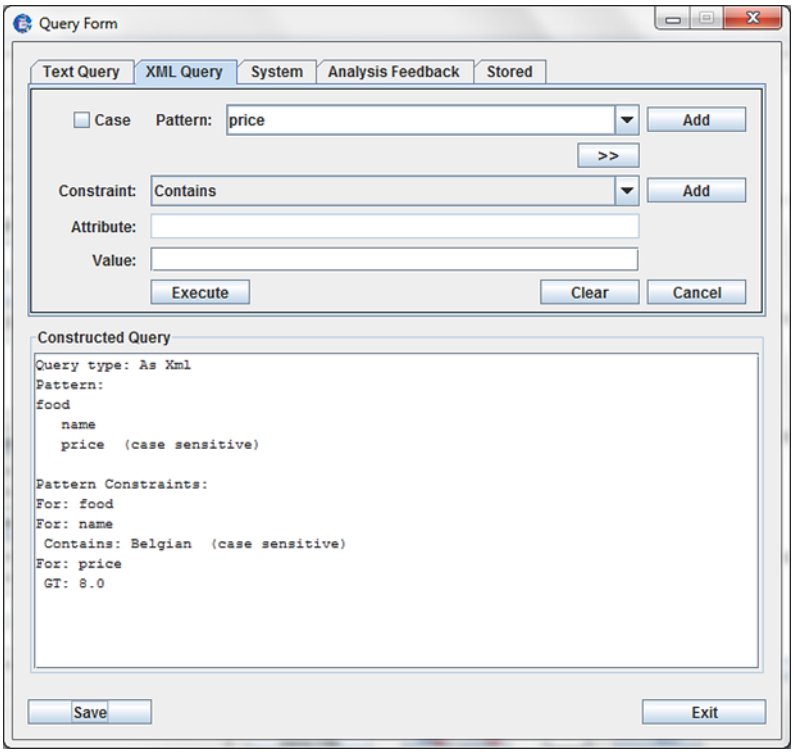


Figure 9. XML-based pattern-style query.

5.2.2.1 Element Pattern

You enter the XML tag names in the `Pattern` field and click the related `Add` button to add them to the query specification. There is a `Case` check box to make the matching process case-sensitive. Wildcards are not considered. The pattern name is added to the same combo box that you enter it in and this also defines its position. Structure is therefore considered. When adding structure, start with the top-most element to be nested and click the right arrow button. This will move the element to be a nested child of the parent element immediately above it. Repeat this for all elements that are to be nested. If there is no structure, the query will try a match to each of the individual elements, when it becomes more like an OR query. If structure is specified, then the process tries to match to the whole nested structure as well. It is also possible to repeat the pattern elements with different sets of conditions, to act again like an OR query, when either set of conditions can be met.

5.2.2.2 Element Constraints

Each pattern element then has a related constraint set that is labelled with the element name as shown. The constraint set can be empty for a match to the element pattern only. If the constraint type is an `Attribute`, it requires the attribute name as well as a value. If it is a comparison, then it only requires the value to compare to. Note that an attribute comparison is an 'equals' comparison only. You can scroll to any pattern name and then add a set of constraints as needed, where the options are as follows:

- **Contains:** if this is selected you also need to enter a text value before adding the constraint. This will then look for text content of the related XML element that contains the text value. The case sensitive option then defines if the match is case sensitive.
- **Not Contains:** if this is selected, the line is not allowed to contain the constraint value.
- **Equals:** if this is selected you also need to enter a text value before adding the constraint. This will then look for text content of the related XML element that matches the text value exactly. The case sensitive option then defines if the match is case sensitive.
- **Not Equals:** if this is selected, the line is not allowed to equal the constraint value.
- **Numerical:** as well as evaluating numerical queries over text lines, it is possible to query specific element text values through a pattern query. If the 'Equals' or 'Not Equals' does not evaluate for true as a text-based comparison, it is checked again as part of a numerical evaluation. As well as that, GT, GE, LT, or LE can also be evaluated as part of a numerical comparison.
- **Attribute:** if this is selected you also need to enter an attribute name and an attribute value before adding the constraint. This will then look for an attribute with the specified name and value that is part of the XML element.

Note: When querying XML, to match the selected elements with the text lines, any additional lines at the start of the XML document (header or comments) need to be removed from the

text that is finally displayed. This is the only minor problem when using this query form. It does not affect the query, but might remove a header line from the XML output.

5.2.3 System Queries

The third query panel can be used to search for files or folders on your whole operating system. These options however allow you to search for more than one term, or to search over the content of certain files.

5.2.3.1 Folder Search

This can perform a general search for folder names. You enter the term that you want to find as part of any folder name. The folder name does not need to be an exact word for word match, but the match is performed only with the last folder name of any path. You then click the `Add` button to add the search to the query. Before it is added, one other option is asked.

- A browser opens for you to select the original folder to search from.

You then click the `Execute` button to execute the query. Any folders that contain the search term are then listed in the main panel text area. You can then open these through a popup menu option, for example.

5.2.3.2 File Search

This can perform searches over folder contents. Any file name can be searched for, but a search can only read TFF-compatible files, that is, text-based or PDF. The content of something like a Word document could not be read, for example. There are two different search types here: to search over the file names in folders, or the file contents in folders. These can be used as follows:

- **File Name Search:** all file names are retrieved and a check is made if the search term exists anywhere in the file name.
- **File Content Search:** all files that can be read are parsed and a check is made if the search term exists anywhere in the file contents.

If these options are selected, the `Comparison` box becomes editable. You enter the term that you want to find in any document content or name. Each file only needs to contain the specified search term in any text sequence and a match with the file name only is made. You then click the `Add` button to add the search to the query. Before it is added, two other options are asked.

- First, a browser opens for you to select the folder to search from.

- Second, you are asked if sub-folders should also be searched.

You then click the `Execute` button to execute the query. Any files that contain the search term are then listed in the main panel text area. You can then open these through a popup menu option, for example.

5.2.4 Analysis Feedback

The fourth query panel can be used to feed analysis results back into the main document.

5.2.4.1 Popular Words

If the analysis has returned a list of popular words, this option will display them in the query form. There are different ways to search for words in a document, but this provides a pre-defined set of words to look for. You can select more than one word, where the query will highlight any line that contains any of the selected words.

5.2.4.2 Highlighting

Highlighting can be used to feed the results of an analysis back into the main text, to highlight the lines selected by the analysis. With this option, you must have performed the analysis operation first (see section 11). If you then select the 'Highlight' query type, the list of word sequences relating to the analysed line numbers are retrieved and you can select which list of lines to highlight. Executing the query will then do that.

5.2.5 Stored Queries

The final panel allows you to load in a stored query. You cannot then easily change it through the GUI form, but you should be able to execute it again. The other parts of the form are not updated or changed, so a whole query model is loaded in and then executed upon request.

6 Selecting Text Areas to Process

It is possible to select areas of text to process, instead of the whole text document. The text is stored in a table format that allows you to select specific rows for reformatting or filtering. This means that you can select an area of the text document and process it with one set of instructions; then select another area and process it with a different set of instructions. When you go to process your text, if an area is selected, you will be asked if you want to process just that area or the whole document, as shown in the dialog box of Figure 10.

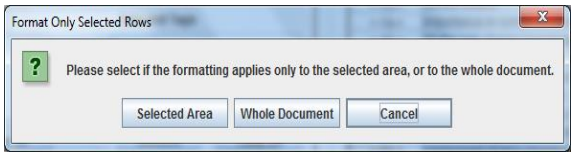
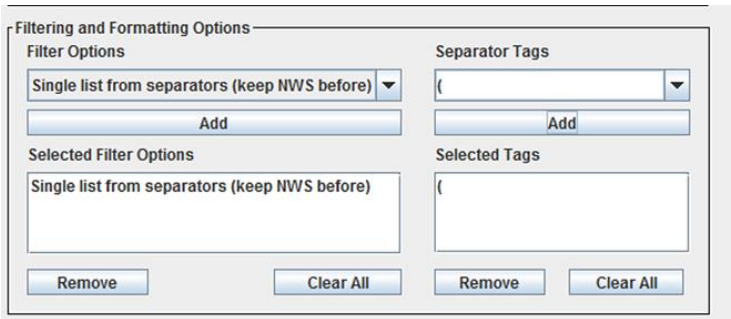


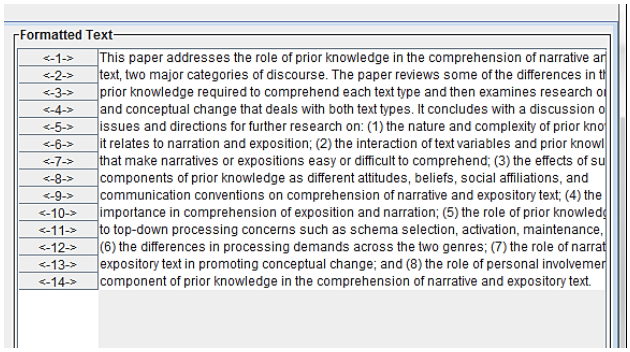
Figure 10. Dialog box giving the choice of what to process.

Figure 11 gives one example of the formatting operators at work. While it requires a number of specific operations, it is still quite arbitrary:

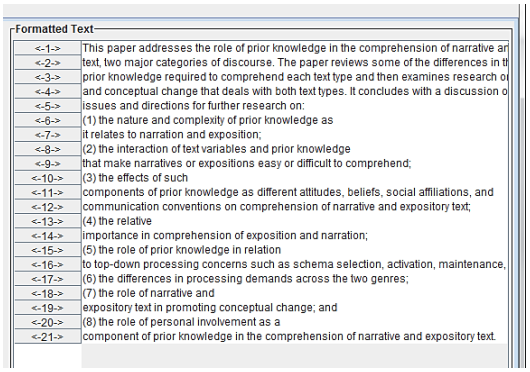
1. A piece of expository text has been loaded into the application as shown in figure (a).
2. The points are then placed on new lines by creating a single column list based on the ‘(‘ separator character, but keeping the character as part of the text and the new line created before the specified separators (Keep NWS before), as shown in figure (b).
3. The points then need reformatting again to the specified width. The text area that covers the points is selected, and the ‘Reformat to new width (include separators)’ option is chosen. This is shown in figure (c).
4. When you click to reformat the text again, if you choose to reformat just the selected text (Selected Area), then the reformatting is performed as shown in figure (d).



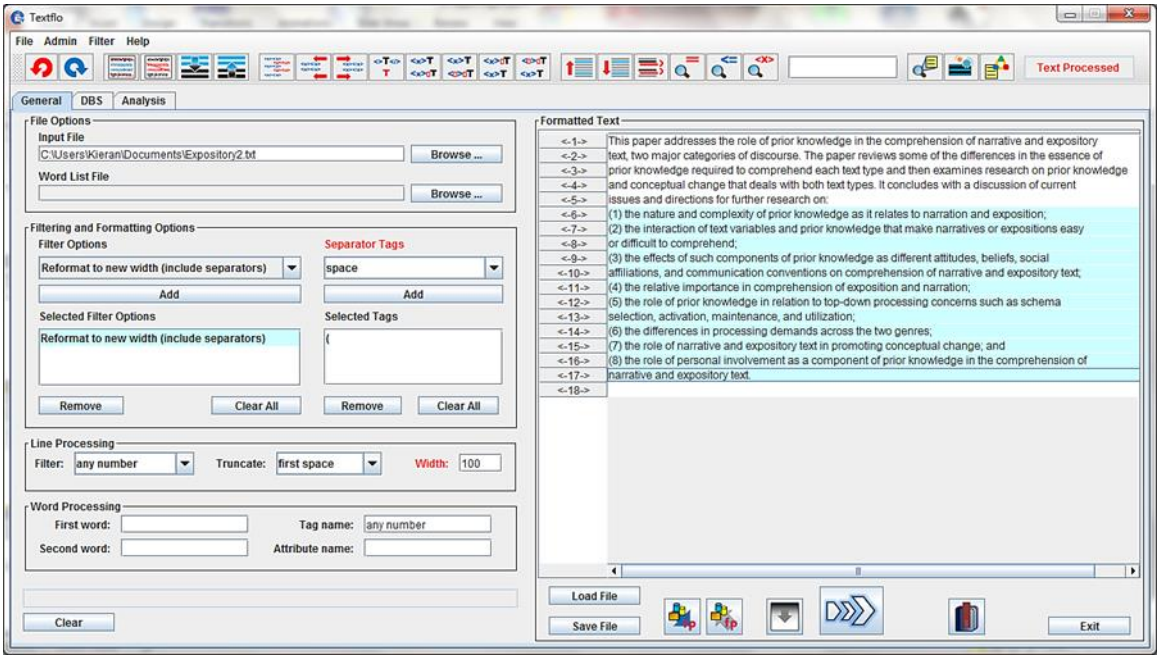
Formatting option to convert from a to b.



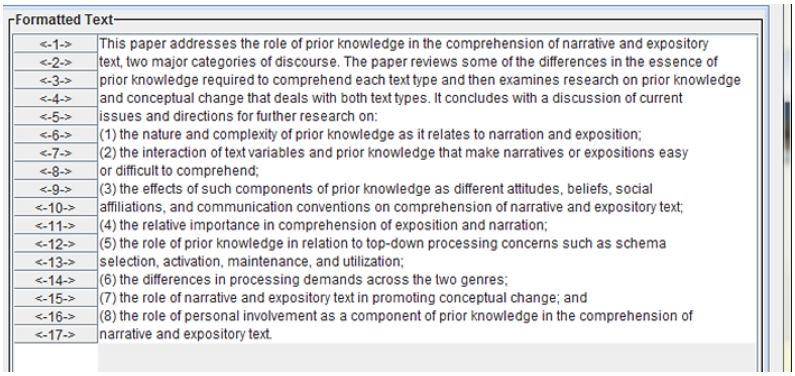
(a) Before formatting



(b) After formatting



(c) Formatting option to change width, but keep '(' as new line start



(d)

Figure 11. Sequence of operations to reformat a piece of text to a specified width, while placing certain sections on new lines as well.

7 Project and Temporary File Analysis Form

This is a new feature that is still being updated. The form now contains 3 panels, relating to popular files and text lines, with a small amount of formatting analysis. The first panel is built similar to the organiser, where you store snippets of text instead. The second panel can display a set of recent files or references that you have looked at, opened, added, or whatever. The third panel can suggest lines in your current document. These options are still a work-in-progress, but they might be useful as is, as they provide another view over your data.

7.1 Project Panel

This is similar to the Organise, where you can open a new project and add another single level of hashtags or categories. For each category, you can paste in text snippets that can be from analysed text, or form any other document on your computer. This is shown in Figure 12. You copy the text into the top text area and enter an optional reference. You can then `Add` the text snippet to the category group. This will allow you to create lists of text snippets, categorised over whatever is relevant to the project.

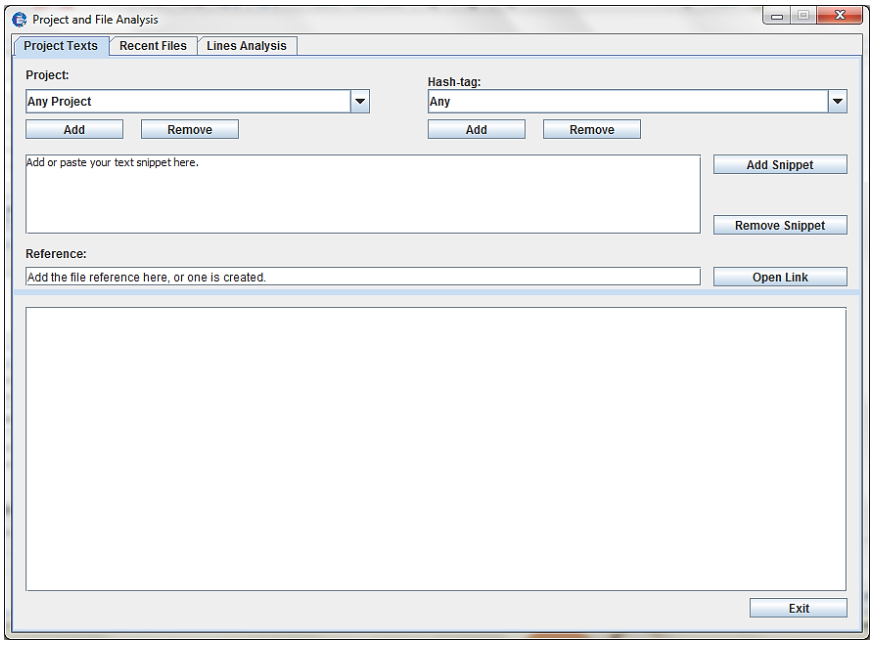


Figure 12. The Project form.

To use this form, you can create a new `project` and then add `hashtag` categories to it. The default `Any Project` or `Any` categories cannot be deleted, but anything else can. You can copy text to the clipboard from any external processor, or from the main GUI text view, using the `Copy to Clipboard` popup menu option. Note that you have to specifically select, or highlight, the text section to copy first. You then paste the text into the top text area. It also has a paste popup menu option. You can also add a reference link, or one is created if you do not. If you then click the `Add Snippet` button, the text and reference key is added to the snippets category. If you select a snippet, the related key should be displayed as well. The key needs to be unique, but if you repeat one, a number gets added to the end. Before trying to open the link, simply remove the number part first.

7.2 Temporary or Recent File References

The second tab in the form lists a number of recent files that you have looked at. The most recent should be at the top, as shown in Figure 13. Files get added to the list through a number of different processes, but mainly to do with opening or loading in a new file or reference. The organiser references can also be added if the organiser is opened from the main GUI – that is – it has a reference to the main GUI itself.

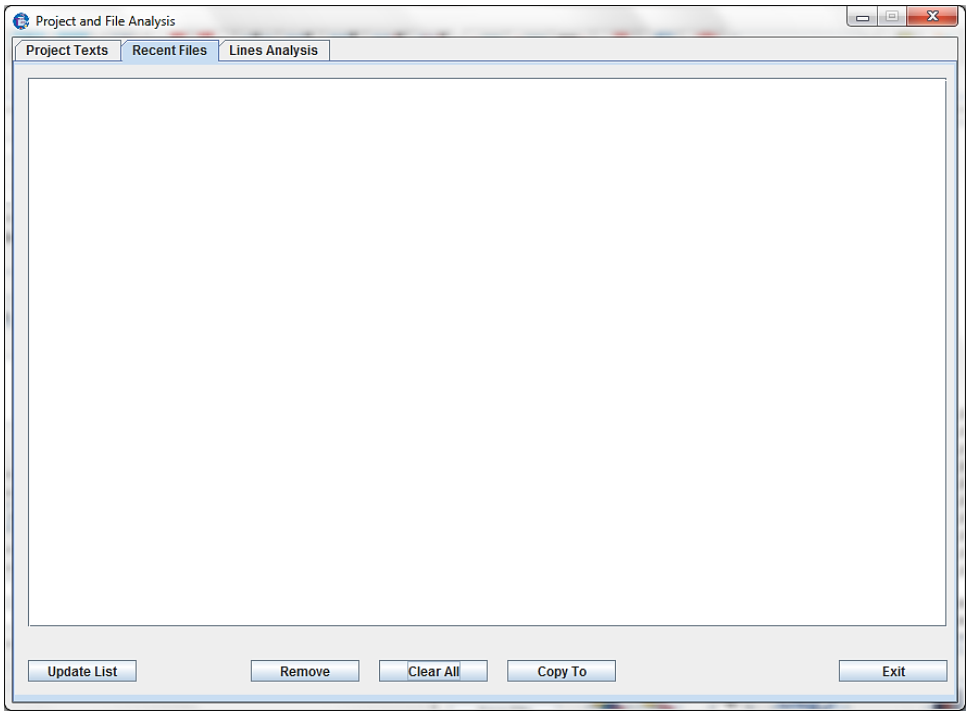


Figure 13. Temporary or Recent File References form.

From this form, you can select a reference or group of references and click `Remove` to remove them. The `Clear All` button will remove the whole current reference list. The `Copy To` button allows you to copy the details to the main GUI text output area. This will overwrite any text currently being displayed, but it will keep the text as a list of file paths. You can then open the Organiser and copy them there through its `Refs from Main` button. The `Exit` button exits the form.

7.3 Line Suggestions

There is also a line suggestions form that can be opened through the `Filter – Temp Analysis File` menu option or the toolbar. It performs some very basic comparisons and suggests lines that occur more commonly. It might be useful for recognising header or footer lines, for example. Figure 14 shows this form.

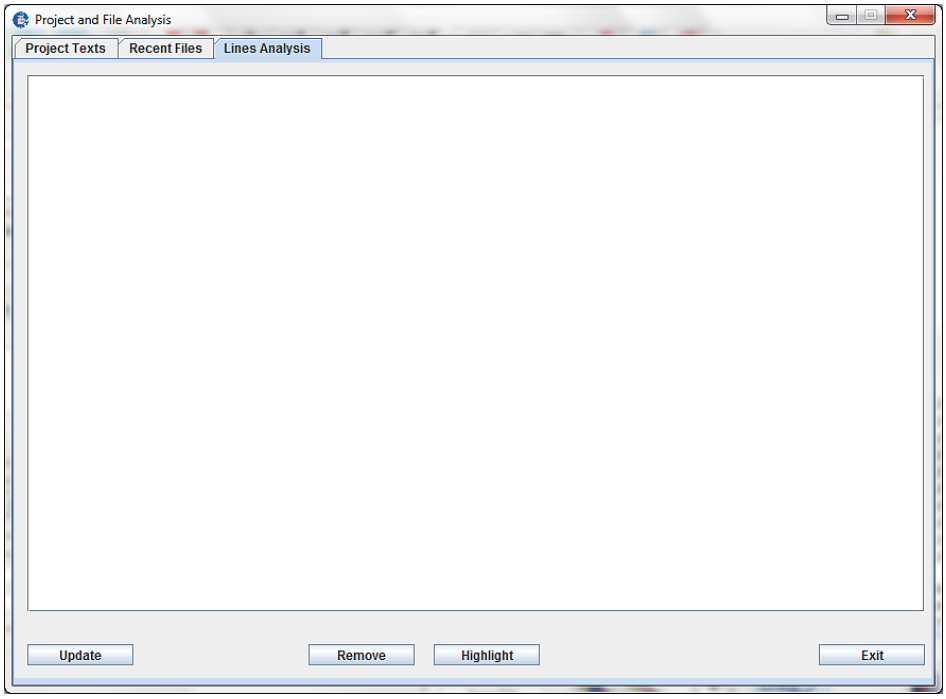


Figure 14. Line Suggestions form.

It is not very scientific, but it will list popular whole lines and also popular starts or ends (2 words) of lines. You can select any of these from the list and then either 'highlight' them in the main text or 'delete' them completely. For documents with very few pages, it might not make any suggestions, but the re-formatting should be easier there.

8 Document Organiser

The application also comes with a document organiser that can be used to group related text document references and links, or with the PDF reader; a library of papers or other documents, can be grouped together based on a list of categories, a free text description and a list of keywords. The document file paths or references can be listed under a set of these values, allowing you to see what each document relates to. It is also possible to search over the group categories and find related documents through these searches. The organiser can also be opened by itself, as a separate application and is shown in Figure 15. It also has its own user guide that is installed with the main application, or is downloadable from the web site at http://distributedcomputingsystems.co.uk/Documents/tffOrganiser.pdf. See the separate user guide for details on how to use the organiser application.

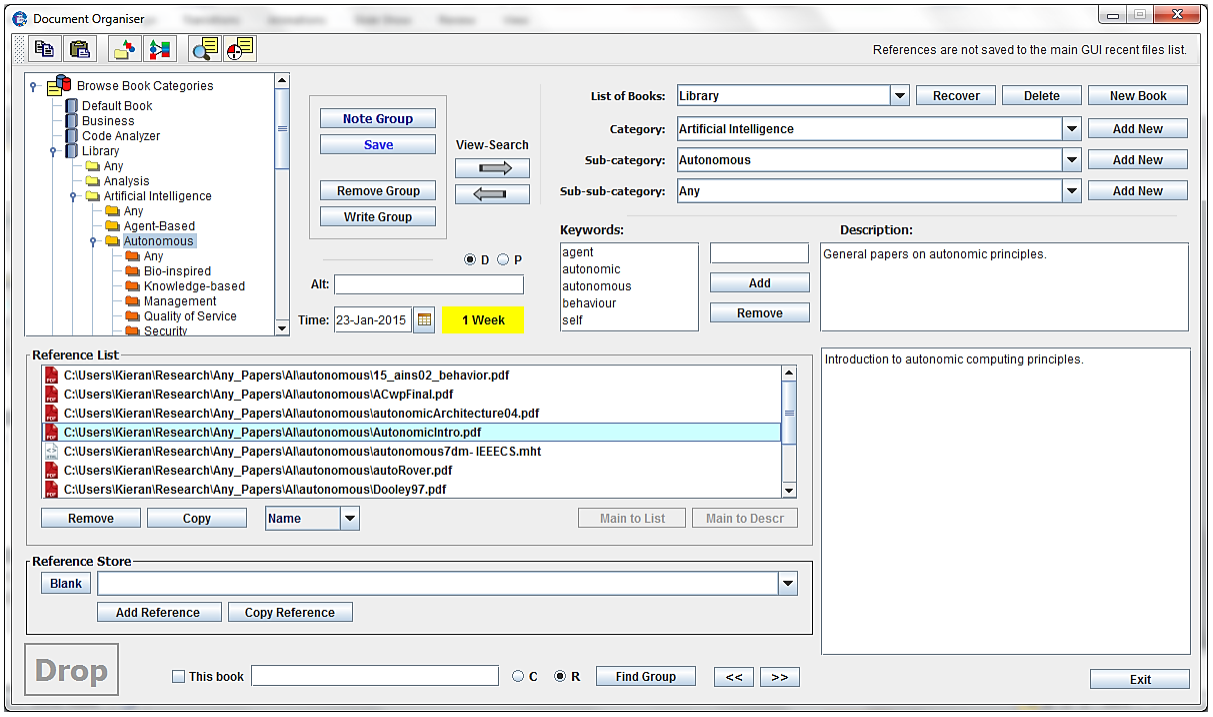


Figure 15. Document Organiser form.

9 Filtering Example

The following example shows one filtering process on the file ‘menu.xml’ that is included in the zip download. This is just an illustrative example of how easily the text can be changed by a number of operations. Figure 16 is the application GUI again, but with a number of filtering options entered as follows:

1. The file is the XML menu file and so the first option is to remove the XML tags (toolbar button or filter option).
2. A set of filtering options are then run in sequence:
   2.1. The common words in the word list are removed.
   2.2. A single column list of words is then created from the separator characters that have been entered as the separator tags.
   2.3. This list is then reordered into a nested ascending order. Note this this now needs to be done through the BDS panel.

This part of the operation could be saved as a stored filer procedure, but the sorts are now separate and part of the DBS operations. Figure 16 shows a stored procedure that has been loaded in and executed, to create a single list of words or terms.

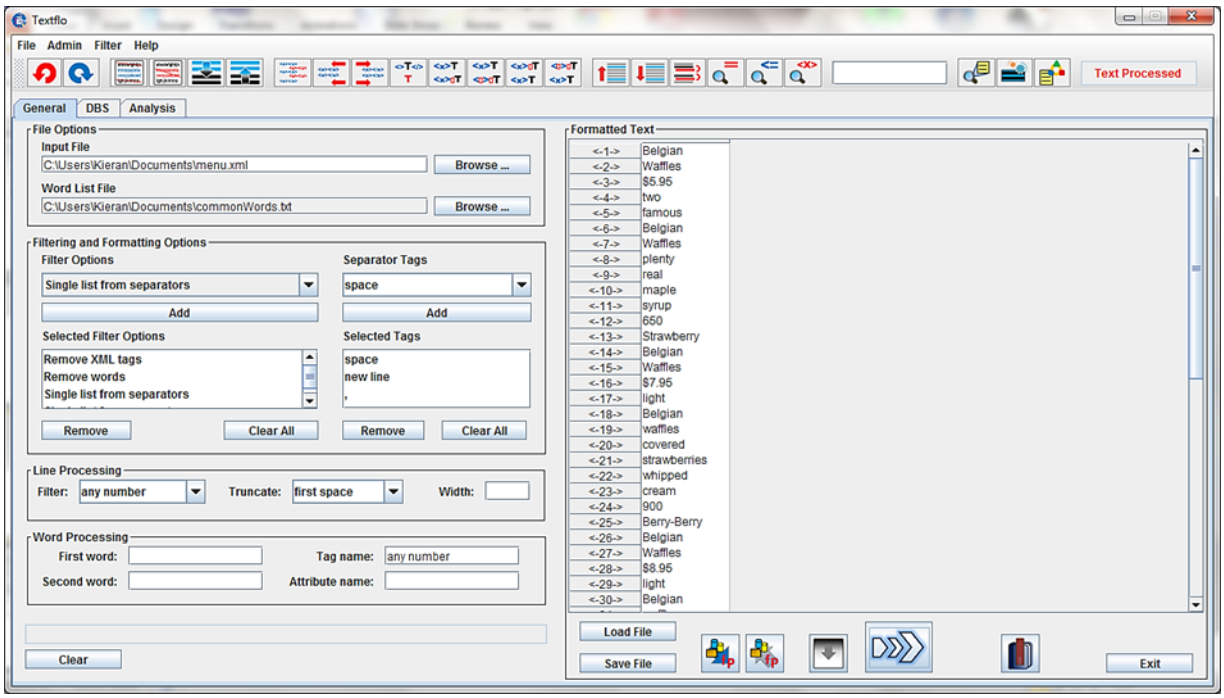


Figure 16. GUI displaying filtering options and formatted text.

The filtering process is performed by clicking the `Format File` button (the large sideways arrow). The sorts have now been moved to the DBS panel because they probably relate more closely with tabular or more structured data. This is described in more detail later (section 10), but Figure 17 briefly shows the sort type that was selected and the result of the sort on the list of terms, displayed in the grid view. Note the entry for a column number, to allow you to select a specific column to sort over.

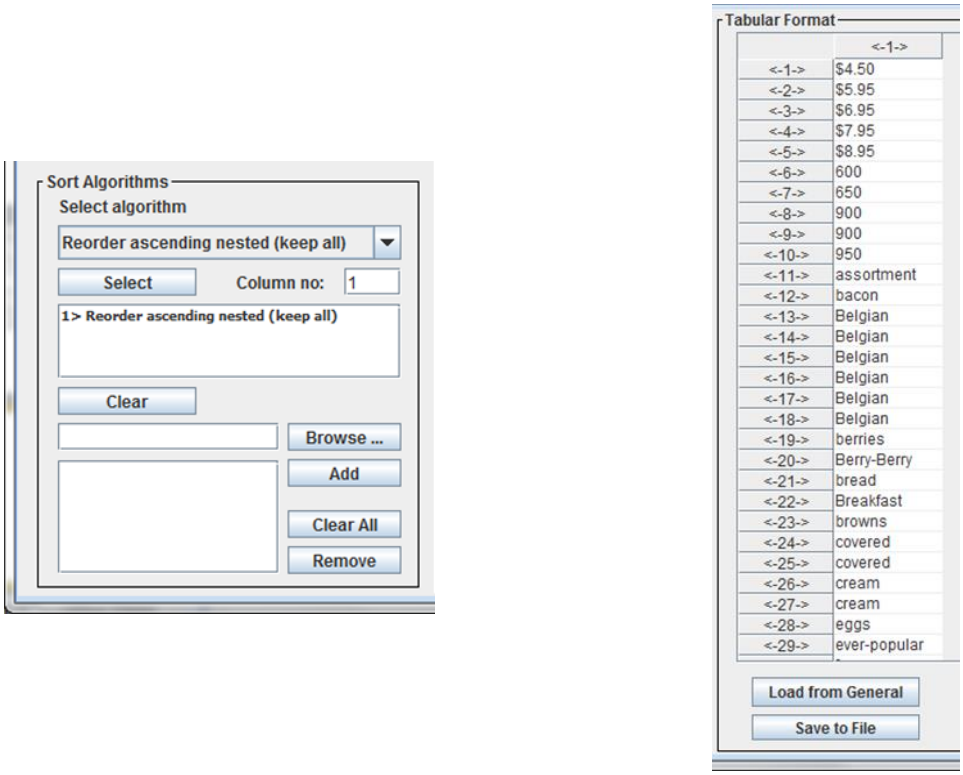


Figure 17. Single column of terms sorted into ascending nested order.

You can see how the first group starts with numbers and orders ascending up to ‘thick’, before the second group starts, etc. For the specified document, this was probably not a useful operation, other than to show how easily the text can be transformed into a different format. In general, it is not entirely clear where this might be useful, but it is a form of text or data mining and would be much more useful over structured or tabular data. You can also manually set the patterns to look for (the nesting order) using the `Ordering and XML` options, when the text will be changed according to that rule instead.

10 Database and Sorts

This panel has been updated into one that could be compared to a very simple database, due to its tabular format. You have the option of viewing real database tables in the main GUI, or creating your own tabular data. Any tables that you create can be re-formatted and changed. The grid view is also generally transferred to the main output area, which might provide a slightly different format to it. You can see the text in this tabular format in Figure 18, where the data contains both words and numbers. The general formatting options would, for example, have problems removing the first column of numbers. You can therefore use the DBS panel to load the text into a grid or tabular structure, where each individual tokenized element is separate and can be removed independently of any other element.

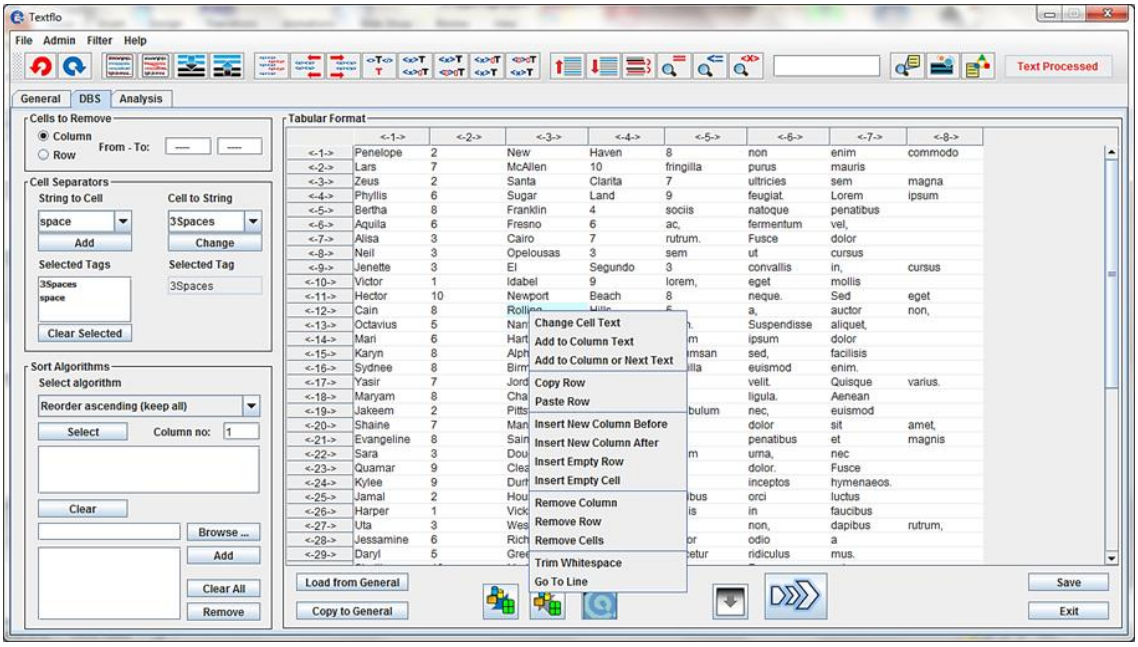


Figure 18. DBS panel with data loaded into a tabular format.

10.1 Load and Save Options

You would typically load in the text from the main panel, using the `Load from General` button. If you then want to save re-formatted text, you can click the `Save` button. This will open in the general documents area, allowing you to select any folder or file name to save to. The `Copy to General` button perform a direct copy of the grid to the main text output area, without additional formatting. It performs the same function as the save button of the main panel. With a tabular format in mind, there is also a dedicated 'db' folder that is created inside the base 'tffData' folder when the application is installed. If you click the `Load`

from DB button, it will automatically open at this location. The Save to DB button will also automatically open at this location. You can then browse to somewhere else, but this folder is provided just as a convenience.

When loading in data, you can specify a set of separator characters that should be used to separate text for each column. The default is the set of whitespace, but other characters can be included. When saving, the default is again a space, but something like a comma could be used instead. When converting from the text to the grid, these values can be set in the Selected Tags list in the Cell Separators group box. The separators for the input are on the left and the separator for the output is on the right. Only one output separator is allowed, so if you update it, it will change that single value. Note that there is now also a Clear button that will clear the contents of the grid only and not the main text area.

10.2 Cell-Level Processing

The General panel allows you to choose a sequence of operations that are automatically applied to the whole document. The DBS panel provides additional functionality through a grid structure that allows you to select specific columns and rows. This means that you can manually select a certain column of data and remove it, without it being part of any particular general filtering rule or condition. So the manual panel allows processing of the text down to a single word or cell, at any position and therefore would allow any kind of text removal filtering to be possible. The graphic of Figure 18 shows what the manual filtering panel looks like. The graphic of Figure 19 firstly shows some random XML data that is then converted into tokenized text through some basic instructions, to be displayed in the grid. The left-hand graphic shows the data in XML format as a series of records. The XML tags are removed, where the right-hand graphic shows the same data with only the XML content kept. To convert from the left-hand format into the right-hand one, the following operations can be tried. This could then be saved as a filter procedure, for example:

1. Replace all ‘<Name>’ words with something such as ‘<Name>–‘. The text, even if it is XML, can still be treated as just ordinary text. The ‘–‘ character can then act as a marker.
2. Remove all XML Tags using either the toolbar or filter list option.
3. Create a single line of text from the filtering options.
4. Tokenize this line, or create a ‘single list from separators’, where the separator is defined only to be the ‘–‘ character that was added.

To load the text document that is currently displayed in the General panel Output area, you click the Load from General button in the Manual Panel. This loads in any text that is displayed in that text area. If this text area is empty, then the program will try to load the data in from the file path.

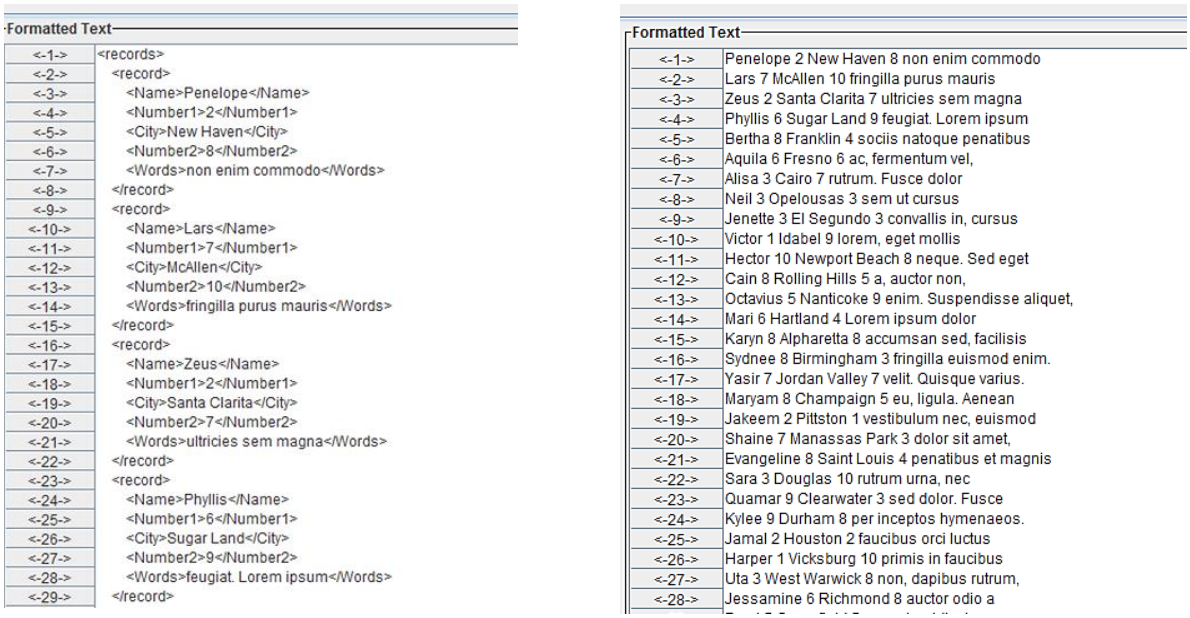


Figure 19. Data in XML format converted to text.

10.3 Manual Filtering Options

The first row and the first column in the grid give each row or column a numerical index value to identify it. These numerical indexes are surrounded by brackets, to clearly define them.

For example '<- 1 ->' is a row or column indexer.

By default the grid structure is generated by separating the text using the whitespace characters – that is – newline, space(s) and tab. The `Separators` group box allows you to re-generate the grid using a different set of separators that you would specify there. This is the same process as for the `General` panel separators options.

The left-hand side of the panel also has a number of filtering options. These options relate to whole columns or lines. The `Cells to Remove` group box stores options to remove whole columns or lines. The top combo box is the column or row to start removing from and the bottom combo box is the column or row to stop removing at. The `from` word identifies the combo boxes to start removing from and the `to` word identifies the combo boxes to stop removing at. You specify the start and end columns or rows by manually entering a numerical value to represent that column or row.

To remove the selected rows or columns, you then press the large `Arrow` (reformat text) button. This will remove the selected cells, update the grid table and then also update the text output area in the `General` panel. If you decide that you do not want to keep this filtering,

you have the option to revert back to the previous text content through the undo/redo buttons. This now resets any highlighted line numbers, so you will not be able to redo or undo highlighting as well.

10.4 Popup Menu

The grid table also allows a popup menu to appear that provides additional formatting options. This is shown in Figure 18 where a number of the cells have been selected and the popup menu item is showing. Note that the cells need to be highlighted first before they can be selected. So you need to click on the cell first to highlight it and then right-click to open the popup-menu. The popup menu then has a number of functions that can be performed on the grid text. These are as follows:

1. **Copy Row:** This copies the contents of the currently selected row. Note that a cell must be highlighted for a row to be selected.
2. **Paste Row:** This pastes the contents of the copied row into the position of the currently selected row. The selected row is moved down one position to allow this.
3. **Change Cell Text:** This option allows you to enter new text that completely replaces the text in the selected cell.
4. **Add to Column Text**: Allows you to add new text to an existing column. You are firstly prompted to enter the text that you want to add. You are allowed to add it before or after the existing column text. The existing text is represented by the property `%CURRENT_TEXT%` and could be different for each row in the column. You can select whether to add the new text before or after the existing text and with or without a space in-between. After your selection, the column is updated to the new text value.
5. **Add to Column or Next Text**: Allows you to add new text to an existing column or the next one encountered with the specified value. In this case, only cells with the specified value (case sensitive) are changed and if the value is not found, no cell in the row is changed. So a search is performed from the specified column onwards and the first cell that matches the entered value is changed only. You are firstly prompted to enter the value for the cell that you want to change. You are then prompted to enter the text that you want to add and then you are allowed to choose whether to add it before or after the existing text. The rest of the process is as in option 1.
6. **Insert New Column**: Allows you to insert a new column into the grid. This can be an empty column with a special empty cell tag, as described next, or you can insert a new column that is filled with a particular word or group of words. This is similar to the previous option, but because a new separate column is inserted, it is reversible. There are options to insert either before the selected column or after it. You are prompted to enter the words to insert. If you cancel this or leave it blank, you can then enter an empty column if you wish.
7. **Insert Empty Row / Insert Empty Cell:** Because the conversion to a grid format removes any empty lines in the text, you will automatically lose that formatting. Two other options allow you to insert empty spaces back into the document. One option from

the popup menu allows you to reinsert blank lines back into the grid structure. This is done through the `Insert Empty Row` menu option. This will allow you to reinsert the paragraph formatting. The blank lines are represented in the grid with the special word sequence:

```
<-- tff-empty row -->
```

However, if you view the actual text in the `General` tab, you will see that this sequence is converted simply into a blank line. So if you then save the text, you will be saving blank lines, or the format as shown in the `General` pane's text output view. Another similar option allows you to insert blank cells into the grid structure. This is done through the `Insert Empty Cell` menu option. The empty cells are represented in the grid with the special word sequence:

```
<-- tff-empty cell -->
```

This is useful if you want to delete a column from the table, but you want to keep one row intact, even with the word that it contains in the specified column. You can then insert an empty cell in that row and column position, and when the column is then deleted, the empty cell will be removed instead.

8. **Remove Column:** Allows you to remove the selected column from the grid.
9. **Remove Row:** Allows you to remove the selected row from the grid.
10. **Remove Cells:** Allows you to remove the selected cells from the grid. This means that only partial rows or columns can also manually be removed, keeping the rest of the text the same.
11. **Trim Whitespace:** Allows you to trim any extra whitespace characters from a list of words in a particular column. If it happens that when you copy text from some source, the formatting of the original text has left unwanted whitespace characters in some position in the text, you can tokenise this into the grid and then remove the whitespace through this option.
12. **Go to line:** This does not require a line to be selected first and will scroll the text to the entered line number

10.5 HyperSQL Database Manager

The HyperSQL database manager has been added as an option, with a button beside the local database buttons at the bottom of the tab. Clicking on the button opens up the manager GUI interface, exactly as has been provided by the package. This is an interface to different types of database, both local and online, where to use it you will need to read the documentation at the web site (http://www.hsqldb.org/). The database manager itself has not been changed, the only difference is the fact that a returned SQL query is presented in the DB grid and not in the HyperSQL GUI itself. If you make any changes to it in the Textflo form, this will not

change the database tables in any way, so it is only the view of an SQL reply that you can use. Updating database tables through HyperSQL will not read any of the re-formatted Textflo text, for example. HyperSQL provides other modes of remote connection, including servlets that are not part of Textflo. You will need to read their documentation to find out about that. There are some other forms and functions that are part of the HyperSQL GUI that are not directly related to textflo, but would help with the database management. You can read their documentation about that.

The GUI interface, shown in Figure 20, is the default one provided by the HyperSQL team. If you try to create a standalone database, the path should default to a 'hsql' folder in your 'db' folder, in the default `tffData` location. So you can create a connection and then execute SQL queries using the second form, on the database tables. The returned result set is displayed in the Textflo DB panel and will include each column name. If you then read from the main panel again, the column names are lost however.

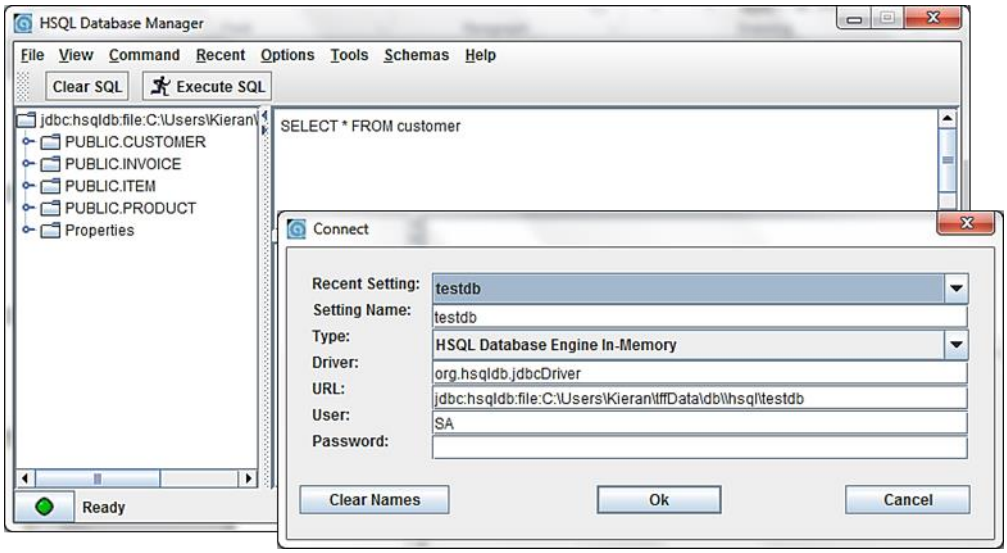


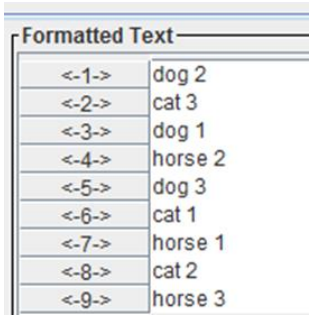
Figure 20. HyperSQL Manager GUI

10.5.1 3Spaces Separator

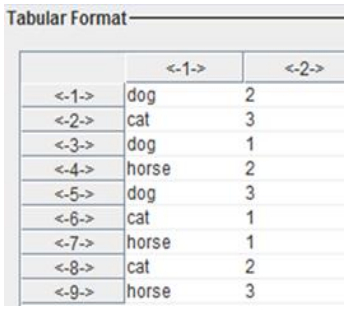
There is an additional separator type for the DB panel, called `3Spaces`. The tabular format is also represented in the main output text area and so, as it is repeated there, it might be useful to provide a slightly different view of it. The default setup is to choose 3Spaces as the separator. This means that if you copy from the main text area to the grid format using this separator, it will separate columns on 3 spaces and not just 1 space, allowing for the original columnar format to be preserved. A rule-of-thumb might be to have 'String to Cell' and 'Cell to String' use the same separator(s), but also that it is distinct.

10.6 Word Sorts

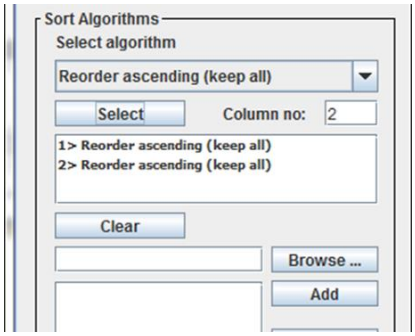
This is a new and useful addition to the DBS panel. The sort options have been moved from the main panel to here. These are displayed in the `Sort Algorithms` group area. A sort would typically be associated more with a single list of terms and the grid structure allows this to be performed over different sets of terms. Therefore, when adding a sort option, you can also specify a column that the sort is to be run over. There is also the text box at the bottom, where you can manually specify your own word order for the 'to list' sorts. These filters are now quite useful, as you can perform one sort inside of another one and select more specifically what data to sort over. After more than one sort, the groups might become fragmented, but that could be data-specific. While the data type cannot be specified, some effort is made to converting to numbers, if the whole column can be used that way. So sorting numbers is automatically carried out, at least in part.



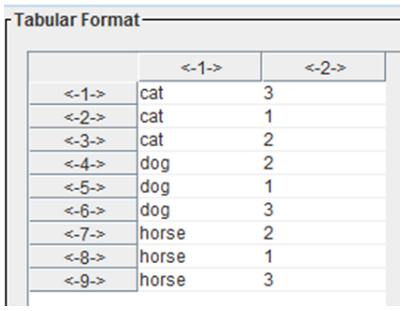
15a – text file



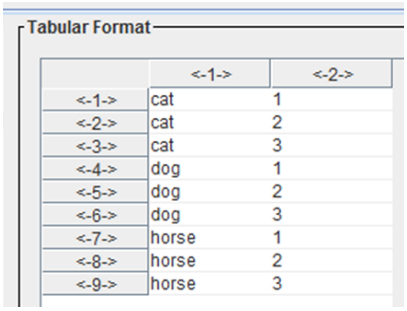
15b – DBS grid



15c – two sorts



15d - just sort 1



15e - sorts 1 and 2

Figure 21. Sort plus nested sort, both ascending, on two columns of data.

The process is as follows:

1. A text file has been loaded into the main panel (15a) and then loaded into the DBS panel (15b).
2. Two sorts have been selected (15c). Both sorts are ascending, but the first is on column 1 and the second is on column 2. The column number is displayed first, followed by the sort

option. Note however that the second sort is not over the whole dataset, but over each sorted section of the first sort.

3. The first sort operation would produce the re-ordering of the lines, shown in 15d. The ‘cat’ words are first, then the ‘dog’ words and then the ‘horse’ words. The second sort would then take each group – cat, dog or horse – and re-order column two into ascending for that group only. It should therefore re-order the numbers for each group, as shown in 15e.

11 Analysis

The application allows for a limited amount of analysis through an Analysis panel. The analysis is based mainly on word counts and comparisons; but it could be accurate enough to give a useful assessment of how similar two sets of text or word lists are. The default options include the standard line, word and character counts that you would find in a Word Processor. Options also exist however to remove or change some word/number combinations, as the text file is pre-processed or filtered. The analysis options that you can select from then include word or word sequence frequencies and also clustering or comparison evaluations that use more sophisticated algorithms. Figure 22 shows what the analysis panel looks like. This analysis has been carried out on the menu document again, with the XML tags removed. As you execute an analysis, a number of message boxes will help to make sure that you are analysing what you mean to. When you have confirmed each option, there is a final message box with a full description of the analysis.

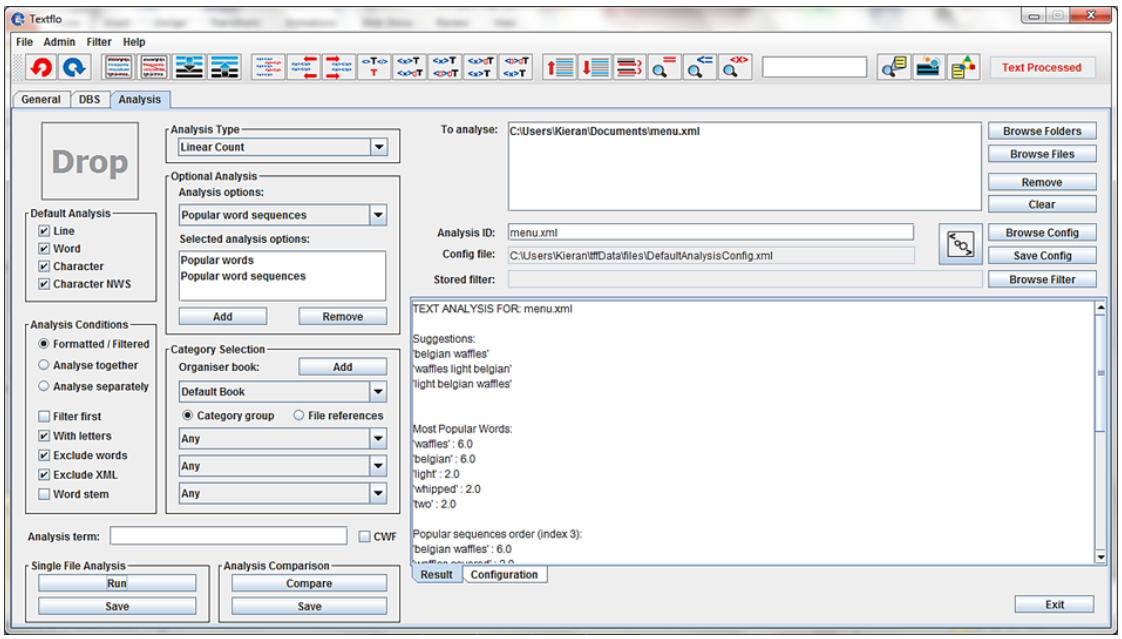


Figure 22: Analysis Panel with the statistical analysis of the displayed text.

11.1 Configuring the Analysis Process

Some of the analysis options are yes/no, or on/off options. These are provided by the set of check boxes on the left-hand side. Other options are more specific to the particular analysis that is to be performed and may require an actual value as well. These can be entered through a Config form and also saved or retrieved from an XML file. The default configuration settings file is saved in the ‘tffData/config’ folder and is loaded during the application startup.

If you look at the Configuration text area tab instead of the Result text area, you will see what the configuration parameters are. If you click on the Config icon button, a form will open that allows you to enter the configuration for the selected analysis algorithm. The form will also enable or disable the values that are relevant to the algorithm. You can enter the values into the form and then, if you click the form's Add button, the details are written to the Configuration text area. Note that the config description now contains an entry describing exactly what analysis type it belongs to. This must match the type of analysis being carried out, even if the other values are OK. The 'Configuration' text area is also editable and so you can make manual changes to any of the values if you wish, but not the XML tag names. You need to make sure that you keep the XML format of the text correct at all times, as this text will subsequently be read and parsed in that format only. The related Browse Config button allows you to load in a different config form, while the Save Config button saves it to a file.

11.2 Analysis Type

The Analysis Type group box at the top of the panel allows you to select the type of analysis to carry out. The system currently comes with basic linear counts, clustering based on related words, or clustering of whole documents based on previous analyses. The linear counts can give a list of the most popular words or word sequences. For a single document, the clustering can give lists of lines that are most similar in the selected text. This can then be fed back into the main panel to highlight sections of the text. Different clustering algorithms can be used to compare a set of previously analysed documents, or file reference groups retrieved from the organiser, as described later.

11.3 Analysis Options

The analysis options that are available depend on what analysis type has been selected. They are declared in the Optional Analysis area, where you select from the combo box there and then Add the option to the list. You can also select an option on the list and click the Remove button to remove it again. To refine the analysis further, there are a set of check boxes on the left-hand side that can be used to alter the text before it is analysed. Basically, you can select from these options and get some sort of result at the end of it. Some options will analyses all of the documents as a single piece of text and some will analyse them separately. There is also an output of the selected options with their meaning before the analysis is run. The radio button options apply only to the single file analyses, not the analysis comparisons. A summary of the check box options is as follows:

1. **Formatted / filtered text:** If this is selected, then only the text in the main GUI window will be analysed. You can therefore filter or change the text content first, before analysing it. You cannot add this changed text to a file list however and so to analyse with other texts, you would need to save it and then list the file path instead. To do some pre-formatting to all texts, you can select from the filtering options for any scenario. If this

option is not selected, then the list of `To Analyse` file paths are read and their texts processed as follows:

2. **Analyse together:** If this is selected, all files in the file list will be read and analysed together as a single text analysis.
3. **Analyse separately:** If this is selected, it forces each text in the file list to be analysed as a single separate document, with the result then saved to a file. The saved analysis file is assigned the name of the original file plus an `.anls` file type extension. So without this option all files are treated as a single group, but with this option, each file is analysed separately, which makes batch processing of several files easier.
4. **Filter first:** If this is selected, a saved filter procedure is used to process every text file first, before the other options are applied. The stored filter procedure can be browsed for and the file path added to the `Stored filter` text area. Stored filter procedure files have a `.fpr` file extension. So both the file and this check box option need to be specified. Then, each file in the list is read and processed by the filter procedure first. The resulting text is then further processed by the analysis options, before the analysis result is returned.
5. **With letters:** If this box is selected, then each word that is considered must contain at least one letter. So valid numbers would not be considered.
6. **Exclude words:** If this box is selected, then the terms in a common word list are removed from the text first, before analysing the remaining words. There is a default common words list that is saved in the `tffData/files` folder. You can also enter a word list into the main GUI panel just blow the text file path. If a word list file path is entered there, it will be used instead.
7. **Exclude XML:** If this box is selected, then the XML tags are removed from any XML document, before the remaining text is analysed. Note that this can change the line number settings, as the analysed text is different to the text that is displayed.
8. **Word stem:** If this option is selected, then word stemming is applied, to try to group the same word with different endings together. For example, ‘word’ and ‘words’ would be considered to be the same. This can only be used with the English language, as the stem roots are only known for this language.

11.3.1 Further Selection Options

Some additional search or analysis options. There is also a final description of all of the analysis choices that have been selected. You can therefore check everything again and confirm, or cancel if it is not correct.

11.3.1.1 Search Term

There is an `Analysis term` text field, where you can enter a term to search for. The analysis will then only include results that contain that term. If you are looking for something specific this will be a quicker option, but it does not apply to every analysis type. Some operations will take the whole text field as a single search term. Others will allow several

terms, separated by commas ‘,’. If the option allows only a single term, all commas are removed, but if multiple terms are allowed, then you can have more than one word in a term as well.

11.3.1.2 Compare with first File Only

Another option here is to click the `CWF` check box (compare with first). If selected, all analyses will be compared to the analysis of the first file in the list only. You can still find this information if analysing between all files, but it might give a slightly clearer output.

11.4 Text Content and File Lists

The analysis can be performed on existing formatted/filtered text, or on text currently stored in files. You can also pre-filter the text to be analysed by selecting or de-selecting the check box options of the `Format Conditions` group box, as described in section 11.3. Any file path specified in the `General` tab’s `Input File` box is initially displayed at the top of this tab in the `File List` area. In addition to this:

- You can use the `Browse Folders` option of the `File List` area to load in all file paths from the directory that you select. This also allows you to load in all file paths from all of its sub-folders.
- You can use the `Browse Files` option to load in a set of file paths that you specify.
- You can select category groups or reference lists from the organiser groups, see also section 11.8.
- You can add a file path by dragging it to the `Drop` zone.

You can then perform an analysis that considers the text in all of the specified files, either as a single analysis or as a separate batch process.

11.5 Saving or Retrieving Analyses

There is now a dedicated `anls` folder in the default ‘tffData’ folder, where analyses are automatically saved to. If, for example, you select to analyse the files separately and save to result files, a folder is created at this location and the new analysis file set is saved there. If you browse to select files, it will automatically open at this location. To save a single or combined analysis requires saving the current text description. The `Save Analysis` button allows you to save this result description in XML format.

If you enter a name in the `Analysis ID` text field, this is used as the analysis identifier for all files analysed during the operation. If the analyses are subsequently compared, the file name should be used to identify each one instead. If this field is empty, then the filename of the first listed file is used as the group identifier.

11.6 Analysis of Individual Files or File Groups

The files to be analysed are typically read from the file list in the `Analysis` tab. The only exception to this is if there is an existing filtered or formatted document and the `Formatted / Filtered` check box is selected. In that case, the analysis applies to the filtered text instead. If a list of files are specified, if any are recognised as special files (previous analysis type or category group type) then they are removed before the analysis process starts. Only raw text files can be analysed this way. These files can however be included in a comparison or clustering operation.

11.7 Comparison Analyses

It is also possible to compare the analyses of files. This helps to determine how similar the content of the files are. A list of files to compare must be entered and would typically include files of type analysis (.anls), but also raw text files or descriptions of category groups. The parser will recognise these different types and convert them first, before performing the analysis comparison. The program will also try to advise on what is being analysed and under what conditions.

- If you reference existing analysis files, then they are read as is. The analysis type for the comparison is taken to be the one currently selected in the `Analysis Type` combo box. Any analysis files should have been created previously from that type only.
- You can also reference raw text files. In that case, they will be converted into analysis models first, based on the selected algorithm and options, before being compared to the referenced analysis files. Each raw text file will produce a new analysis model.
- You can also reference file lists stored in your organiser. The `Category Selection` area of the tab allows you to browse through your saved organiser categories, to add a group. If this is added to the analysis process, the file list relating to the category group is retrieved and analysed as a single group. This produces a single analysis model that is then compared with the other ones. This is a useful way to determine what category or group a new file might belong to, as part of a clustering process. Note that you can now select from different books in the same operation, where the first element of the display path is the book name.

The comparison analysis produces a comparison for every analysis file against every other one.

11.8 Category Selection of Organiser Groups or Files

The `Category Selection` area of the panel allows you to browse through your saved Organiser categories, to add a group's details. This is the default setting, but you can switch to a `File References` setting using the radio buttons, to load in the file paths instead. Loading the file list is useful, because the browse options will browse to files that are all in

the same folder, whereas a category group might store selected files from different folder locations. Each file can then be compared with every other file that is listed, whereas a group is taken to be a single entity. However, if deciding that a new document belongs to any existing groups, then the group option would be correct. Note that comparing arbitrary text documents might not produce large scores of similarity and so some manual interpretation might be required, where relative values are more important than actual ones.

11.9 Analysis Algorithms

A detailed description of exactly what each analysis option does is described in the analysis guide document that is downloadable from the main web site at http://distributedcomputingsystems.co.uk/Documents/tffTextAnalysis.pdf. A text analysis guide has been written to reduce the size of this document and to store the more technical information separately. So that you have a general idea of what each algorithm does, a brief summary follows. There is no hard rule as to what the best algorithm might be and so you will probably decide this based on the results that they provide for the analysis problem that you enter. Some of the algorithms also now have a 'Suggestions' section at the start. If there is a particular result that is repeated or notably better, it is put into a suggestions section for your attention.

11.9.1 Linear Count

This performs basic word counts. It counts the most popular words and also word sequences. This can process a single file or a list of files. A list of texts will be combined into a single document before being analysed. If the 'analyse separately' option is selected, then this overrides the order to combine the text documents and they are analysed separately and saved to individual files. You can also select to look for a specific word or term, through the `Analysis term` field. If this field is used, the analysis will only consider word combinations that include the term and it will also reduce the processing time.

11.9.1.1 Linear Count Comparisons

This option also allows for analysis comparisons, as described in the text analysis document. In addition to that, it will perform a basic word or sequence count over the documents instead of the text content. So if the word or word sequence occurs in a document, it has a value incremented by 1. This is then displayed along with the comparisons of the frequencies generated for each individual document. The suggestions are based only on the frequency percentage values however.

Note: The analysis will currently only add a sequence if the frequency count is larger than 1. It will also include any sequence that contains a smaller one, without incrementing the max sequences count, so the max sequences count value relates to new sequences, where the

output list can be larger and have sequence parts that are repeated. This is just to add some more variability.

11.9.2 Line Cluster

This looks for popular word sequences in the text and returns the lines that they occur on. To view in the application, this can only process one text document at a time, as it is the matching line numbers that are returned and also what you would use next. You can then highlight sets of clustered lines in the main text through the query form. If a file list is specified, only the first file will be considered. If the 'Analyse separately' option is selected however, then the text documents are analysed individually and saved to files. This option might not be 100% reliable with all texts, especially when empty lines are involved and so checking the line numbers would be a good idea. You can also select to look for a specific word or term, through the `Analysis term` field. This type will then only consider word combinations that include the term and it will also reduce the processing time. The conditions of the previous section also apply here.

11.9.3 Clustering Algorithms

There are some clustering algorithms that can be used to 'compare' document analysis results. They implement a number of well-known metrics that compare the similarity or difference between pairs of analyses:

- Cosine similarity and Jaccard coefficient measure the same sort of thing. They measure a set similarity that does not consider the exact placement or the terms.
- The Similarity function is a simple count of the number of terms that are the same. It does not consider word frequency.
- The CF Inverse Doc Freq can also be used for word frequency comparisons and is also the basis for the other metrics. It creates the word list that the other metrics use.
- Kullback-Leibler is a more information-oriented and probabilistic method.

The CF Inverse Doc Freq performs primarily a popular word count. If only one document is analysed, this can produce a similar score to the linear word count. The algorithm looks for the most popular words in a document, but also considers if it is popular in other documents as well. It looks for the most distinguishing features in the document, giving higher scores to features that are popular in one document, but not every document. A text about 'computer hardware', listed with texts about 'computer software', for example, should rank a word like 'CPU' highly, because the software texts would not include it. Note that common words might still rank highly in any text, because they are so numerous, and so they can be filtered out first using the pre-processing options. This option does not use a specific analysis term, as it only considers single words for its clustering.

You can select any or all of these metrics, where each will measure the similarity between the document analyses. The result outputs each heuristic for each document pair in a row. The suggestions section can then try to filter this further. These algorithms can therefore analyse a list of files, as well as a single text document. The options to analyse separately however, will force the analysis of each text individually and then save the result to a file. The analysis process will not change the contents of any files that are listed and so it is advisable to try different scenarios and options, to see what sort of analysis results are produced.

Note: The different metrics can produce different scores, sometimes just in terms of magnitude, so it would be a matter of using the ones that are most appropriate to you. It might also be the case that using a single clustering type is better than combining the results of more than one, especially if they produce different results. So do not assume that adding more options will produce better clusters.

11.9.4 Information Retrieval (Professional version only)

For a comparison analysis, this option will calculate the precision of each of the search terms that are entered, compared to the whole list of input files. For a single analysis, it will calculate a count of each search term in each document and output the totals. It gives some idea of how relevant the search term is to the file group as a whole.

Acknowledgements

This software product uses the HyperSQL database manager (http://www.hsqldb.org/). The calendar date GUI component is provided by Microba (MichaelBaranov.com) and the GUI balloon components are provided by BalloonTip (http://java.net/projects/balloontip/). PDF to Text conversion is carried out using JPod from Intarsys Consulting GmbH (http://opensource.intarsys.de/home/en/index.php?n=OpenSource.JPod). The Microsoft Word processing uses the Apache POI packages (http://poi.apache.org). The common words list has been taken from the Text Fixer web site (http://www.textfixer.com/). Thanks also to TeamBox (http://teambox.com/) for the free file type icon set used by the bookmarks.

12 Appendix A - Filter Options

This appendix describes all of the available filtering options with respect to what the input and output should be for each one.

12.1 Basic Formatting

These options are for the basic reformatting of text as a whole document. They deal primarily with processing each line with the same set of instructions.

12.1.1 Trim Whitespace

Trims any leading or trailing whitespace from every line:
Name: **Trim whitespace**
Input: current text.
Output: current text with all trailing and leading whitespace removed. Blank lines are kept.

12.1.2 Single spaces

Convert the document so that there are only single spaces between each word:
Name: **Single spaces**
Input: current text.
Output: current text with only one space between each word.

12.1.3 Reformat the line width with no other separators

Reformat the text to have the specified line width:
Name: **Reformat to new width (no other separators)**
Input: current text, the maximum allowed line width.
Output: current text converted into a single paragraph, with lines of the maximum specified width.

12.1.4 Reformat the line width and include other separators

Reformat the text to have the specified line width. If however, there is a list of other separator tags, then these can be used to create new lines as well:
Name: **Reformat to new width (include separators)**
Input: current text, the maximum allowed line width.
Output: current text converted into a single paragraph, with lines of the maximum specified width. Lines of shorter length are also possible of a separator tag is encountered.

12.1.5 Replace Word1 with Word2

This simply replaces all of the occurrences of the first word by the second word:
Name: **Replace word1 with word2**
Input: current text, word1, word2.
Output: current text with all word1 converted into word2.

12.1.6 Truncate, keep after a specified character or word

Reformat the text to truncate all lines at the specified character or word:
Name: **Truncate, keep after first character**
Input: current text, the truncating word or character.
Output: current text with each line truncated by removing everything up to this word or character (exclusive) if it is present, or the whole line otherwise.

12.1.7 Truncate, keep after, with a specified character or word

Reformat the text to truncate all lines at the specified character or word:
Name: **Truncate, keep after, with first character**
Input: current text, the truncating word or character.
Output: current text with each line truncated by removing everything up to this word or character (inclusive) if it is present, or the whole line otherwise.

12.1.8 Truncate, keep to a specified character or word

Reformat the text to truncate all lines at the specified character or word:
Name: **Truncate, keep to first character**
Input: current text, the truncating word or character.
Output: current text with each line truncated by removing everything after this word or character (inclusive) if it is present, or the whole line otherwise.

12.1.9 Truncate, keep to, with a specified character or word

Reformat the text to truncate all lines at the specified character or word:
Name: **Truncate, keep to, with first character**
Input: current text, the truncating word or character.
Output: current text with each line truncated by removing everything after this word or character (exclusive) if it is present, or the whole line otherwise.

12.1.10 Text to upper case

This simply converts all of the text to upper case:
Name: **To upper case**
Input: current text.

Output: current text with all characters converted to upper case.

12.1.11 Text to lower case

This simply converts all of the text to lower case:
Name: **To lower case**
Input: current text.
Output: current text with all characters converted to lower case.

12.1.12 Reformat to a single line of text

This simply converts all of the text back into a single line:
Name: **Single line**
Input: current text.
Output: current text with all newline characters converted into spaces.

12.2 Search

These options relate to searching over single lines of text. Case is generally not considered when comparing words, so capitals or upper case is usually ignored.

12.2.1 Remove all lines that contain exactly any of the words in the word file from the text

Remove all lines that contain any of the words in the list from the text. The words must be whole individual words in the line. The word list can be replaced with a single entry in the Filter box:
Name: **Remove lines with (exactly)**
Input1: current text, list of words to remove.
Output1: current text with all lines that contain any of the words in the list removed.
Input2: a single word sequence in the Filter box.
Output2: current text with all lines that contain the word sequence, as a whole sequence, removed. If one of the general options is selected – any letters, any characters, or any symbols – then lines that contain only the general specification are removed. For example, only numbers will remove all lines that contain only numbers.

12.2.2 Remove all lines that contain in sequence any of the words in the word file from the text

Remove all lines that contain any of the words in the list from the text. The words can be part of any text sequence in the line to remove:
Name: **Remove lines with (contains)**

Input1: current text, list of words to remove.
Output1: current text with all lines that contain any of the words in the list removed.
Input2: a single word sequence in the Filter box.
Output2: current text with all lines that contain the word sequence, as part of any sequence, removed. If one of the general options is selected – any letters, any characters, or any symbols – then lines that contain these in any sequence are removed.

12.2.3 Remove all lines that start with the filter text

Remove all lines that start with the specified filter text, as specified by the Filter box:
Name: **Remove lines that start with**
Input1: current text, list of words to remove.
Output1: current text with only the lines that start with any of the words in the list removed.
Input2: a single word sequence in the Filter box.
Output2: current text with all lines that start with the word sequence, as a whole sequence, removed.

12.2.4 Keep only the lines that contain exactly any of the words in the word file from the text

Keep only the lines that contain any of the words in the list from the text. The words must be whole individual words in the line to keep:
Name: **Keep lines with (exactly)**
Input1: current text, list of words to keep.
Output1: current text with only the lines that contain any of the words in the list kept.
Input2: a single word sequence in the Filter box.
Output2: current text with all lines that contain the word sequence, as a whole sequence, kept. If one of the general options is selected – any letters, any characters, or any symbols – then lines that contain only the general specification are kept. For example, only numbers will keep all lines that contain only numbers.

12.2.5 Keep only the lines that contain in sequence any of the words in the word file from the text

Keep only the lines that contain any of the words in the list from the text. The words must be part of any text sequence in the line to keep:
Name: **Keep lines with (contains)**
Input1: current text, list of words to keep.
Output1: current text with only the lines that contain any of the words in the list kept.
Input2: a single word sequence in the Filter box.
Output2: current text with all lines that contain the word sequence, as part of any sequence, kept. If one of the general options is selected – any letters, any characters, or any symbols – then lines that contain these in any sequence are kept.

12.2.6 Keep all lines that start with the filter text

Keep all lines that start with the specified filter text, as specified by the Filter box:
Name: **Keep lines that start with**
Input1: current text, list of words to keep.
Output1: current text with only the lines that start with any of the words in the list kept.
Input2: a single word sequence in the Filter box.
Output2: current text with all lines that start with the word sequence, as a whole sequence, kept.

12.3 Words and Lines

These options relate to processing the text as separate words or lines. Each word or line can be selected or filtered independently.

12.3.1 Remove all separator tags

Remove all of the separator tags from the text:
Name: **Remove separators**
Input: current text, list of separator characters.
Output: current text with separators replaced by single spaces.

12.3.2 Remove all lines that have a width smaller than the width specified

Remove all lines that have a width smaller than the specified width. To be used with caution as the ends of paragraphs might also include one or two words:
Name: **Remove lines smaller than**
Input: the current text, minimum allowed width for a line in terms of number of characters.
Output: current text with all lines that have fewer characters than the width specified removed.

12.3.3 Remove all lines that are blank/empty or only have whitespace

Remove all lines that are empty/blank, or only have whitespace:
Name: **Remove blank lines**
Input: the current text.
Output: current text with all blank or empty lines removed.

12.3.4 Remove all lines that are blank/empty or only have whitespace, if there is more than one in a row

Remove all lines that are empty/blank, or only have whitespace:
Name: **Remove blank lines > 1**
Input: the current text.
Output: current text with a maximum of only 1 blank line between text paragraphs.

12.3.5 Remove the words in the word file from the text

Remove all words in the list from the text:
Name: **Remove words**
Input: current text, list of words to remove.
Output: current text with the words in the list removed.

12.3.6 Keep only the words in the word file in the text

Keep only the words in the list in the text:
Name: **Keep only words**
Input: current text, list of words to keep.
Output: current text with only the words in the list kept.

12.3.7 Remove duplicate lines

Remove duplicate lines from the text – that is – only allow one instance of each line, but exactly as it is printed. This is case and space sensitive:
Name: **Remove duplicate lines**
Input: current text.
Output: current text with only one instance of each line.

12.3.8 Remove duplicate words

Remove duplicate words from the text – that is – only allow one instance of each word. This is not case sensitive:
Name: **Remove duplicate words**
Input: current text.
Output: current text with only one instance of each word.

12.3.9 Remove duplicate words in sequence

Remove duplicate words from a text sequence – that is – only allow one instance of each word. A word is removed if it is the same as the previous word. This is not case sensitive:
Name: **Remove duplicate words in sequence**
Input: current text.
Output: current text with only one instance of each word in the sequence.

12.4 XML-Based

These options relate to processing XML text specifically. They should be used along with the toolbar options that can separate the text content from the XML tags.

12.4.1 Remove tags and keep content

Keep only the content of the XML elements. Also a toolbar option:
Name: **Remove XML Tags**
Input: current text.
Output: only the text content of the XML document.

12.4.2 Separate whole tags from text

Place a space between the tag and the text content of any element. Also a toolbar option:
Name: **Separate XML tags**
Input: current text.
Output: the same text, but with a space between each element tag and its text content.

12.4.3 Re-join whole tags with text

Remove the space between the tag and the text content of any element. Also a toolbar option:
Name: **Re-join XML tags**
Input: current text.
Output: the same text, but with the space between each element tag and its text content removed.

12.4.4 Separate tag names from brackets and text

Place a space between the tag bracket and the text content of any element, and also between the tag name and the tag bracket. This allows parsing of the tag name only. Also a toolbar option:
Name: **Separate XML tags to words**
Input: current text.
Output: the same text, but with spaces between each element tag name, brackets and its text content.

12.4.5 Re-join tag names to brackets and text

Remove the spaces between the tag bracket and the text content of any element, and also between the tag name and the tag bracket. Also a toolbar option:
Name: **Re-join words to XML tags**
Input: current text.

Output: the same text, but with the spaces between each element tag name, bracket and its text content removed.

12.4.6 Surround selected section with a tag

Surround a whole section with a single XML element, with the tag name that is specified:
Name: **Surround selected text with tag**
Input: current text, the name of the XML element (Tag).
Output: current text converted into XML, where the whole section has been enclosed in an element with the specified name.

12.4.7 Surround each line with a tag

Surround each line of text with an XML element, with the tag name that is specified:
Name: **Surround each line with tag**
Input: current text, the name of the XML element (Tag).
Output: current text converted into XML, where each line is an element with the specified name.

12.4.8 Surround specific lines with a tag

Surround specific lines in the text with an XML tag, with the tag name that is specified:
Name: **Surround word with tag**
Input: current text, the name of the XML element (Tag), name of the text word, representing a whole line (Word).
Output: current text where only specific words are converted into XML, where each converted line is an element with the specified name.

12.4.9 Convert text to attribute

Convert the text values of certain elements into attribute values:
Name: **Text to attribute**
Input: current text, name of the element (Tag), the name of the attribute (Attribute).
Output: current text where the elements with the specified name have their text values converted into attributes with the specified name. The text value is then left empty.

12.4.10 Remove HTML Formatting

An HTML file can be loaded in as a text file. This can contain a lot of information that is additional to the content that you would read, for example the web page layout or formatting. This option extracts only the html content that is for reading. All of the other text is removed, including all of the other tags. This option is not exact yet, but it is helpful for removing the additional information that might be on the page:

Name: **Remove HTML Formatting**
Input: current text.
Output: current text with only the html reading content remaining.

12.5 Single Lists

These options relate to creating single lists of words from a whole text document. A single list can also mean a sequence of words on each line, if the separating character is not the `space` character.

12.5.1 Single column list

Convert the text into a list with only one word on each line. All whitespaces are replaced with newline characters:
Name: **Single column list**
Input: current text.
Output: current text converted to a list of single words.

12.5.2 Single list from separators

Convert the text to replace all of the separator characters with newlines to generate a single list of words:
Name: **Single list from separators**
Input: current text, list of separator characters.
Output: current text with the separator characters replaced with newline characters.

12.5.3 Single list from separators, but keep non-whitespace separators – new line before

Convert the text to replace all of the separator characters with newlines to generate a single list of words. If any of the specified separator characters are non-whitespace, then they are kept in the formatted text as well. The new line is created before the separator character:
Name: **Single list from separators (keep NWS before)**
Input: current text, list of separator characters.
Output: current text with the separator characters replaced with newline characters and the non-whitespace one kept as well.

12.5.4 Single list from separators, but keep non-whitespace separators – new line after

Convert the text to replace all of the separator characters with newlines to generate a single list of words. If any of the specified separator characters are non-whitespace, then they are kept in the formatted text as well. The new line is created after the separator character:
Name: **Single list from separators (keep NWS after)**
Input: current text, list of separator characters.
Output: current text with the separator characters replaced with newline characters and the non-whitespace one kept as well.

12.5.5 Single list from XML tag names

Parse an XML document to extract only the XML element tag names in order, to generate a single list of words:
Name: **Single list from XML tag names**
Input: current text.
Output: single list of words representing the XML element tag names.

12.6 Reorder the created word list

These options relate to re-ordering lists of words that have been created. The word list can be reordered in several ways depending on the user's spec. The list can be reordered in ascending, descending, or user specified order, and conventional or nested. See the main text for the different options:
Name: **Various – experiment to try them out!**
Input: current text as a list of words, word order (optional).
Output: current text where the word list has been reordered as specified.

13 Appendix B - Default Analysis Configuration File

The analysis configuration file is written in XML format. The default file is loaded into the system at startup from the `config` folder and performs the currently available options of popular word or word sequence counts. The structure of the file is shown in Figure 23:

```
<Analysis_Model>
    <Popular_Words_Number>10</Popular_Words_Number>
    <Min_Nesting_Number>2</Min_Nesting_Number>
    <Max_Nesting_Number>5</Max_Nesting_Number>
    <Sequence_Number>3</Sequence_Number>
    <Min_Word_Length>2</Min_Word_Length>
</Analysis_Model>
```

Figure 23. Default Analysis Configuration File.

The following elements can be configured or changed in the file:

- Popular words number: this is the number of popular words to output. The default value of 10 means that the 10 most popular words will be output with their values. If you change this number then that will change the number that is output.
- Minimum nesting number: This is the smallest number of words in a sequence (consecutive) to measure.
- Maximum nesting number: This is the largest number of words in a sequence (consecutive) to measure.
- Sequence number: This is the number of popular sequences to output for each word sequence number.
- Min word length: Even after removing certain words, the document might still contain words you do not want to count, so this allows you to enter a minimum word size.

So for example, if the minimum number is 2 and the maximum number is 5 and the sequence number is 3, the analysis will output and store the top 3 sequences for 2, 3, 4 and 5 word sequences. It is easy to test or change this to see what it does. The configuration file is editable, so you can change it to whatever you wish and then load/save the new file.